

Show or Suppress? Managing Input Uncertainty in Machine Learning Model Explanations

Danding Wang, Wencan Zhang, Brian Y. Lim

School of Computing, National University of Singapore, Singapore

Abstract

Feature attribution is widely used in interpretable machine learning to explain how influential each measured input feature value is for an output inference. However, measurements can be uncertain, and it is unclear how the awareness of input uncertainty can affect the trust in explanations. We propose and study two approaches to help users to manage their perception of uncertainty in a model explanation: 1) transparently show uncertainty in feature attributions to allow users to reflect on, and 2) suppress attribution to features with uncertain measurements and shift attribution to other features by regularizing with an uncertainty penalty. Through simulation experiments, qualitative interviews, and quantitative user evaluations, we identified the benefits of moderately suppressing attribution uncertainty, and concerns regarding showing attribution uncertainty. This work adds to the understanding of handling and communicating uncertainty for model interpretability.

Keywords: Trust, Uncertainty, Interpretable Machine Learning

1. Introduction

The increasing prevalence of machine learning (ML) has called attention to make it more transparent and trustworthy [55]. The burgeoning research area of Explainable AI (XAI) provides a basis to improve understanding and trust in ML model inferences [2, 16, 18, 55]. Explaining with feature attributions (also called attribution explanations) is one popular technique to

Email addresses: wangdanding@u.nus.edu (Danding Wang), wencanz@u.nus.edu (Wencan Zhang), brianlim@comp.nus.edu.sg (Brian Y. Lim)

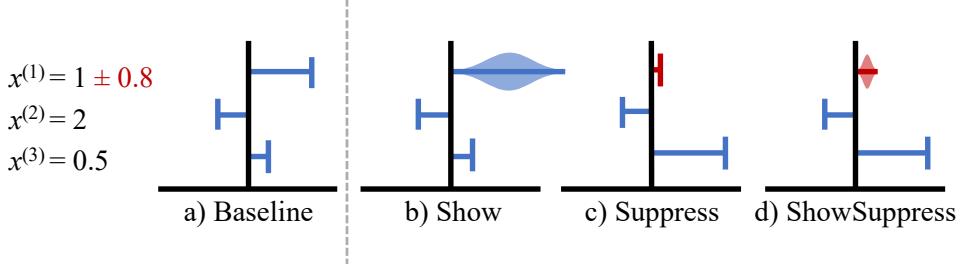


Figure 1: Baseline attribution explanation and three proposed attribution explanations that account for uncertainty. Each horizontal bar in a tornado plot is a feature attribution, and the sum of all bars indicates the total attribution for the model output. Suppose uncertainty in input $x^{(1)}$ is known to be large. This will propagate through the model and can be accounted in its attribution as: a) Baseline explanation which does not handle the uncertainty in $x^{(1)}$; b) Show which visualizes the attribution uncertainty (in this case, with a violin plot); c) Suppress which reduces the attribution to $x^{(1)}$ and reallocates attribution to other features (in this case, $x^{(3)}$); d) ShowSuppress which combines Show and Suppress to suppress attributions towards uncertain features and shows the suppressed attribution uncertainty.

interpret machine learning models (e.g., [4, 57, 63]) by attributing the model’s inference to input features. This indicates whether a feature influences a decision outcome positively or negatively and by how much. An attribution explanation typically involves approximating a linear relationship between input and output, which can also be considered a weighted sum, and is commonly visualized as a tornado plot (see Baseline in Figure 1). Each bar indicates the influence of the feature on the model’s inference. Bars to the right indicate positive influence and bars to the left indicate negative influence; longer bars indicate larger attribution. This intuitive explanation technique illustrates how each feature contributes to the model inference, using basic bar chart literacy.

Attribution explanation techniques focus on conveying salient signals and assume that the input feature values are reliable. In contrast, input measurements are often noisy or uncertain [15], e.g., a temperature reading at a regular time may be affected by a temporary gust of wind, or a sensor may momentarily drop data due to network communication issues. With these fluctuating noisy inputs, the model inference may also fluctuate, and may correspond explanations of these inferences. Thus, attribution explanations should account for and communicate the impact of this uncertainty on users. In this paper, we present two approaches for attribution explanations

to be “uncertainty-aware” to leverage uncertainty information to augment explanation visualization or re-compute the attributions. We are specifically interested in the research questions: In what circumstances will a user trust or not trust an explanation that is based on a set of measurements that are of questionable validity? How can we manage to communicate input uncertainty at inference time to improve trust in explainable AI?

Communicating uncertainty has been an active research topic in supporting people’s use of intelligent systems, such as investigating the impact of showing uncertainty in estimated values [65] in various simpler formats [39], creating novel visualizations for uncertainty [34, 37], and communicating uncertainty in model inference [53, 55]. These approaches demonstrate that showing uncertainty helps to improve user trust in automation and smart systems. We extend the prior research by communicating the uncertainty in the attribution explanation. Specifically, we first propose a Show Uncertainty explanation which propagates the uncertainty of inputs to attribution explanation and visualizes the attribution uncertainty (Figure 1, Show). Interestingly, while showing uncertainty allows users to reflect and calibrate their decisions, some people have an aversion to uncertainty [7, 21, 70, 74], and tend to cope with uncertainty by ignoring, denying, or reducing it by acquiring more information [54]. To support them, we propose an alternative Suppress Uncertainty explanation (Figure 1, Suppress), which reduces the attributions to inputs with high uncertainty to make the explanation less dependent on uncertain information and shift attributions to more certain inputs instead. The interaction design with Suppressed Uncertainty explanations is as follows: i) the user will see that there is uncertainty in some input feature values, ii) be shown an attribution explanation and be informed that the attributions have been suppressed to be less reliant on uncertain features, iii) and be able to toggle the explanation view to see the unsuppressed attribution explanations. From this comparison, the user will learn how attributions of uncertain features have been reduced in magnitude, and partially reallocated to other features. For a wine quality score rating example, due to an input uncertainty of $\pm 1.0\%$ for Alcohol, its attribution is reduced from +2.5 (Figure 8c) to +1.3 (Figure 8b); similarly an input uncertainty of ± 0.2 for Vinegar Taint led to a reduction in its attribution from -1.2 to -0.5; the reduced attributions have been reallocated from -0.4, 0.5, 0.9 to -0.3, 0.7, 1.4 for pH, Sulphates, SO₂, respectively. To suppress uncertainty, we propose two technical methods that regularize the model-agnostic explainer or the predictor model itself. Finally, combining two strategies of showing and

suppressing, we propose the ShowSuppress Uncertainty explanation (Figure 1, ShowSuppress) that first suppresses attribution to uncertain inputs and visualizes any remaining attribution uncertainty. Note that suppressing uncertainty (Suppress) involves shifting the attribution values in the explanation to minimize the influence of inputs with high uncertainty. This does not mean that uncertainty is concealed; the attribution uncertainty can still be shown (ShowSuppress), but they will be smaller than before (Show). This is different from simply not showing attribution uncertainty, which could be the default (Baseline) or with uncertainty suppressed (Suppress).

Given the opposing objectives of these uncertainty management techniques, we investigated the relative benefits and limitations of showing or suppressing uncertainty in explanations compared to baseline attribution explanations. We define a stochastic metric — *Expected Faithfulness Distance* — based on Explanation Faithfulness¹ [63] as a measure of the how poorly (well) an explanation globally agrees with the predictor model when explaining instances with uncertain measurements. In a simulation study, we show how Suppress can reduce Expected Faithfulness Distance. In a qualitative interview study, we learn how participants variously used the different explanation techniques (Baseline, Show, Suppress, ShowSuppress), appreciated the information they provided, but had divergent opinions about their usefulness. In a controlled quantitative user study, we evaluated whether and by how much each explanation technique affected decision quality, confidence, trust in the model predictions, and perceived helpfulness of the explanations. We found that showing uncertainty helps users to understand the system but costs more time, and the decision of users with higher uncertainty tolerance is closer to the system when showing uncertain; and suppressing attribution uncertainty increases explanation faithfulness, user trust, confidence, and decision quality. In summary, our contributions are:

- Three approaches to manage and communicate attribution uncertainty due to input uncertainty: 1) showing attribution uncertainty, 2) suppressing uncertainty by regularizing explainer or predictor models, 3) and a hybrid of showing and suppressing.
- Experiment methods and instruments to evaluate interpretable ML models under uncertainty by employing stochastic metrics, such as ex-

¹Faithfulness is the similarity in predictions between the predictor model and explanation (explainer) model that explains the predictor.

pected faithfulness of hypothetically uncertain instances, and a within-subjects experiment design to evaluate explanation-assisted decision making with input, attribution, and model uncertainties.

- Empirical findings from i) a characterization simulation study, ii) formative qualitative interviews, and iii) a summative quantitative user study to show a) how showing attribution uncertainty improves understanding and trust, but is not helpful to improve decision making, and b) suppressing attribution due to uncertainty improves trust, helpfulness and decision making.

This work adds to the research on explainable AI under uncertainty and highlights the importance to manage and communicate attribution uncertainty to users. From the quantitative and qualitative results of our user studies, we identified the benefits of moderately suppressing attribution uncertainty, and concerns regarding showing attribution uncertainty.

2. Related Work

This work studies the implication of considering uncertainty within explanations of machine learning (ML) models. First, we define the scope of uncertainty that we study, point out that current interpretable ML techniques do not handle uncertainty, and summarize existing methods to handle uncertainty by showing or suppressing it.

There are different concepts and aspects of uncertainty in machine learning: stochasticity and insufficiency in training data [42], possible deviations of a model output [39], probabilistic models with uncertain model parameters [25] and noisy inputs to the model during inference [1]. In this paper, we focus on the last source of uncertainty from noisy input data; this is usually caused by measurement and estimation error. We focus on how input uncertainty should be handled during explaining model inference rather than during explaining model training. Although there are some models that intrinsically modeling or calculating input uncertainty, we focus on more prevalent models that do not model uncertainty and yet face input uncertainty in model inference time. Our paper handles attribution uncertainty due to uncertain input and differentiates from works that identified or handled attribution uncertainty due to the instability of explanation generation process [29, 31, 78, 79].

We note that there may be different reasons for uncertainty in data, such as miscalibrated sensors, lost data connection, adverse or unexpected

environmental changes. Determining these causes is beyond the scope of the machine learning modeling we discuss in this paper.

2.1. Improving Trust With Intelligible, Interpretable Models

The recent interest in explainable AI and interpretable machine learning has given rise to many types of interpretable models and model explainers, e.g. [10, 44, 47, 48, 57, 63]. Typically, attribution explanation generates feature importance by attributing model output to input features. Local Interpretable Model-agnostic Explanation (LIME) [63] is a popular method that generates an instance-level attribution explanation by approximating predictor model locally around the query instance with a linear model as an explainer. Besides the model-agnostic attribution explanation LIME, Integrated Gradients (IG) [72] is an attribution explanation specific to explaining neural networks by integrating the gradient of model output along each input dimension. However, these explanation methods assume that there is no error in the input features and do not consider the attribution uncertainty. Therefore, we propose approaches to providing uncertainty-aware attribution explanations by showing or suppressing strategies. In this paper, for the model-agnostic explanation, we extend LIME by visualizing and regularizing explanation 6 by input uncertainty. To make model-specific attribution explanation uncertainty-aware, we show an example with IG.

The aforementioned works focus on technical innovation to generate explanations, but it is important to ensure that explanations are human-centered to support human reasoning [2, 55, 73], and be integrated into the iterative design pipeline [20]. Many insights can be learned from observing how users use and interact with explanations. Lim and Dey investigated how users interacted with different query types to seek information and learned how little time users spent on consuming explanations [52]. In a design probe study with data scientists, Hohman et al. identified the importance of supporting interaction in explanations [33], and in a lab study, Cheng et al. found that interactive explanations help to improve user understanding [12]. Cai et al. identified what concepts pathologists care in medical images and explains the model by providing similar images based on the similarity of user-defined concepts [8]. Wang et al. explored how tailored explainable AI can help clinicians to improve medical decision making by mitigating specific cognitive biases [73].

In this work, we employ a quantitative empirical controlled user experiments and conduct a qualitative interview to study how users interact with

model explanations under input uncertainty.

2.2. Handling Uncertainty by Showing

Researchers have found that showing inference uncertainty, confidence or accuracy can affect user trust and task performance [3, 31, 55, 65]. Antifakos et al. [3] found that showing uncertainty can decrease task time when uncertainty is low, while Rukzio et al. [65] found that this can hurt task time. Lim and Dey [53] found that low uncertainty can raise user trust, but showing high uncertainty will lower trust, even if the model behavior is correct. Yin et al. [77] found that communicating the model stated and observed accuracy can have a compounded effect on trust.

Many studies communicate uncertainty by showing a numeric score, or ordinal labels (high/low text, traffic light icons, etc.). However, uncertainty can be complex, and several visualization methods have been proposed to display them in interfaces and charts [6, 35, 40, 66]. Hypothetical outcome plots (HOP) help users to understand stochastic events by animating instances on a graph, sampled at random based on its uncertainty distribution [34]. Kay and colleagues found that quantile dot plots help to improve user perception of probabilities and help with decision making [24, 39].

2.3. Handling Uncertainty by Suppression

The aforementioned studies identified cases where showing uncertainty is useful, but can also sometimes be harmful. Indeed, it may be better to suppress uncertainty, since some people have lower tolerances and poorer coping strategies regarding uncertainty [54]. Even domain experts, such as data scientists who are experts in data uncertainty [7] and clinicians regularly make decisions under uncertainty [70] seek to reduce uncertainty. Lipshitz and Strauss [54] identified several uncertainty-coping strategies such as acknowledging, reducing and suppressing. They found that people acknowledge uncertainty, seek more information to reduce uncertainty, and ignore or deny uncertainty. These works suggest that suppressing uncertainty in explanation can better help the users who are not comfortable with uncertainty in decision making. The suppressing technique is inspired by the Dempster–Shafer evidence theory [69]. When sources of evidence (input features) have different reliability (various input uncertainty), to combine the effects of unequal-reliable evidence, there are several methods [46, 68, 76] that discount the effects of unreliable evidence. In the domain of interpretable machine learning, no research has been found to use an uncertainty suppression strategy

to support users who are uncertainty intolerant. Our paper aims to fill this gap by proposing a Suppressed Uncertainty Explanation and evaluating how it affects user trust of machine learning and confidence in decision making.

Some recent researches on explanation robustness relate to our uncertainty suppression approach, but they vary in their scope and objectives. Ghorbi et al. [27] demonstrated that attribution explanations of deep learning are sensitive to adversary perturbation. Handling this, several works aimed to improve the model and explanation robustness. Chen et al. [11] improved robustness by adding a regularization term to the training loss function to penalize the difference between attribution explanations of neighboring instances; this makes explanations of neighboring instances consistent, no sudden changes. Singh et al. [71] improved robustness against adversarial attacks by regularizing their model training loss to penalize differences between attribution explanations of instances and nearby adversarial examples. In contrast, we desensitize the attribution explanation of each instance by adding a regularization term to the training loss function to penalize attributions to input features with high uncertainty; this makes explanations less dependent on features with uncertainty, because attributions to these features will be reduced in magnitude. Furthermore, our aim is not to improve the robustness of a certain type of model, but handling attribution uncertainty to support people who are uncertainty intolerant. Model robustness is a by-product of our Regularized Predictor method. These methods also suppress attribution uncertainty by regularization, but the difference is that they suppress the uncertainty in explanation generation while we suppress the attribution uncertainty due to input uncertainty.

In summary, we focus on the uncertainty in machine learning model explanations due to uncertain inputs, and we aim to extend the research field of human-centered explainable AI by investigating how attribution uncertainty facilitates user understanding of model inference and uncertainty and how attribution uncertainty affects user decision making, trust and confidence. Drawing on and extending prior work from the perception of uncertainty, uncertainty visualization, and regularization in machine learning, we propose two uncertainty handling approaches, showing and suppressing attribution explanation uncertainty, for users with different uncertainty tolerance.

3. Technical Approach

In this section, we introduce how we show attribution explanation uncertainty, henceforth called attribution uncertainty, by sampling hypothetical instances and how we suppress uncertainty by regularization in both model-agnostic and model-specific explanations. For the model-agnostic explanation, we base our two uncertainty-aware explanation techniques on LIME [63], a popular model-agnostic attribution explanation method. This allows post-hoc explanations to be made uncertainty-aware without needing to re-training the predictor model. We first briefly introduce LIME, focusing on how it can provide linear attribution explanations, then describe our method to transparently visualize uncertainty in attributions (Show Uncertainty), and describe how we added a regularization term to penalize on input uncertainty (Suppress Uncertainty). These techniques can be combined to provide the hybrid ShowSuppress explanation. While using model-agnostic explanations support convenient implementation and wider adoption of explanations, making model-specific explanations can further improve explanation faithfulness. Thus, we also propose regularizing the training of the predictor model by input uncertainty to suppress attribution uncertainty, and we show an example of a regularized neural network model explained by Integrated Gradients [72].

3.1. Instance-Based Attribution Explanation

LIME [63] is an instance-based, model-agnostic explainer $g_f(\cdot)$ that explains the inference for each single instance x_0 without needing to know the internal mathematical mechanics of the underlying predictor model $f(\cdot)$. The explainer $g_f(\cdot)$, usually a linear model, is trained by sampling data points in the neighborhood of the query instance x_0 to faithfully describe the inferred outcome from predictor f . Formally, the linear explainer model for LIME is simply defined as

$$g_f(x) = w^\top x = \sum_d w^{(d)} x^{(d)}, \quad (1)$$

where $x^{(d)}$ is the d -th feature value, and $w^{(d)}$ is its linear weight. $w^{(d)}x_0^{(d)}$ is the feature attribution that indicates how influential the d -th feature is for inference on instance x_0 . These feature attributions are commonly visualized in a tornado plot (see Figure 1, Baseline), which are shown to be good for visualizing the influence of multiple variables [23]. In this paper,

we follow the established use [49, 52, 53, 63] of tornado plots for feature attribution explanations. To train the local explainer, a training dataset s_{x_0} is first constructed by sampling in the neighborhood of the query instance x_0 . Neighbors that are closer to x_0 will have a higher influence on the explainer training, encoded as weight c_{x_0} . LIME minimizes the difference between predictor outputs and explainer outputs in the neighborhood subset, which means the linear explainer model should be faithful to the predictor model within the neighborhood of the query instance, and the explainer should have consistent inference with the predictor:

$$\sum_{x \in s_{x_0}} c_{x_0}(x) \cdot (f(x) - g_f(x))^2 + \lambda \|w\|_2^2, \quad (2)$$

where $\lambda \|w\|_2^2$ is the L2-norm regularization on the weights w of the explainer to control for sparsity. Higher λ makes the explainer model sparser with smaller feature attributions. Next, we describe two approaches to manage uncertainty with LIME.

3.2. Showing Uncertainty in Attribution Explanation

Consider feature $x_0^{(d)}$ with input uncertainty $\epsilon^{(d)}$, i.e., $x_0^{(d)} \pm \epsilon^{(d)}$. We aim to propagate this uncertainty to the feature attribution in the linear model explanation, i.e., $w^{(d)} \cdot (x_0^{(d)} \pm \epsilon^{(d)})$. We generate hypothetical instances using Monte-Carlo Sampling on the probability distribution $P(\epsilon)$ of the error ϵ and compute the feature attribution for each hypothetical instance. We repeat this for each feature with uncertainty.

To illustrate the distribution of attributions to the readers, we use violin plots [32] to visualize the distribution of uncertainty in attributions. Figure 1 b) shows an example of Visualized uncertainty as a violin plot for each feature in a tornado plot, specifically, the uncertainty in the attribution for the first feature. Instead of using a T-shape bar to indicate a single-point value for attribution, the violin plot illustrates the probability density of attribution.

3.3. Suppressing Attribution Uncertainty in the Explainer

To reduce the attribution uncertainty, we propose to suppress attribution due to uncertain features in the explainer model. This is not simply concealing the attribution uncertainty but changing the explanation to be less dependent on uncertain features. Specifically, we apply an additional

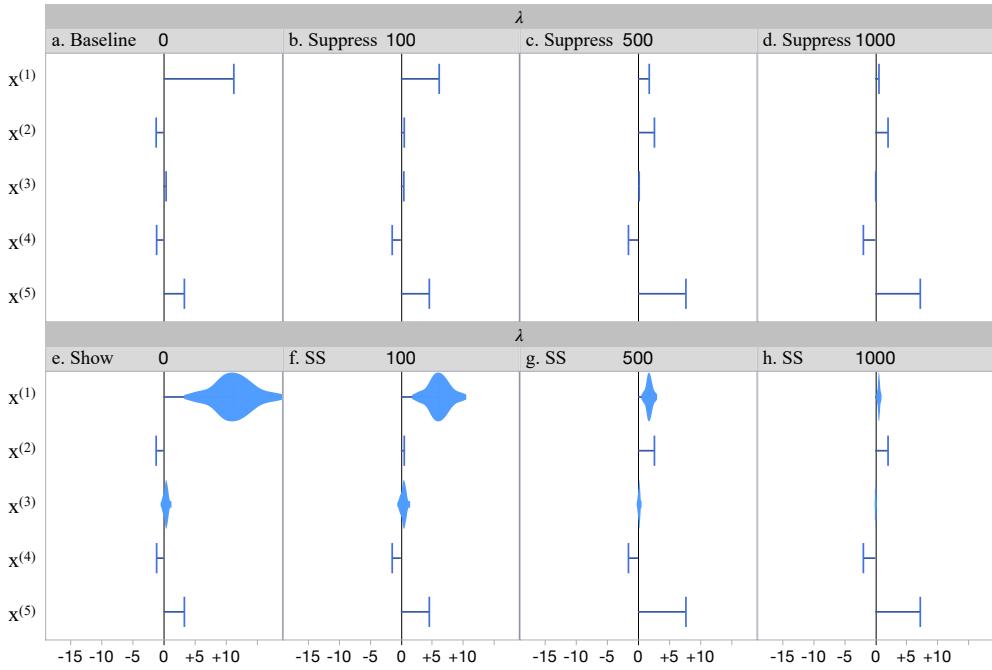


Figure 2: Example of Baseline, Show, Suppress and ShowSuppress (SS) explanations. The attribution explanations in the first row do not show attribution uncertainty while the second row visualizes attribution uncertainty in violin plots. In this case, uncertainty only occurs in features $x^{(1)}$ and $x^{(3)}$, with $x^{(1)}$ having a higher uncertainty. When regularization weight λ is zero, no feature attribution is suppressed and the attribution uncertainty is large (see a.Baseline and e.Show). As λ increases, the attributions to feature $x^{(1)}$ and $x^{(3)}$ are re-attributed to other features (See Suppress explanations in b,c and d). The bottom row ShowSuppress explanations illustrate that the uncertainty in explanation is suppressed when λ increases.

regularization penalty² on the LIME explainer to suppress the attribution uncertainty. This Regularized Explainer $\tilde{g}_f(\cdot)$ decreases the attribution to uncertain features, but this causes a shift to increase attributions to other features to maintain the explainer’s inference faithfulness to predictor output.

We adapt the objective function in Eq 2 for training the explainer to

²Regularization: adding a secondary objective to the training loss function to train the machine learning model so that it performs better with respect to the secondary objective. In our case, we penalize the model from depending too much on inputs with high uncertainty, therefore, we added a loss term regarding input uncertainty.

include penalizing input uncertainty:

$$\sum_{x \in s_{x_0}} c_{x_0}(x) \cdot (f(x) - \tilde{g}_f(x))^2 + \underbrace{\lambda \|\sigma \circ w\|_2^2}_{\text{Attribution Uncertainty penalty}} \quad (3)$$

where $\lambda \|\sigma \circ w\|_2^2$ is the weighted L2-norm regularization term that penalizes weights if there is uncertainty in their corresponding input, σ is the input uncertainty vector whose d -th element is the standard deviation $\sigma^{(d)}$ of the uncertainty $\epsilon^{(d)}$ in $x^{(d)}$, w is the weights vector whose d -th element is the linear weight for element $x^{(d)}$, and \circ is the Hadamard product operator. Any larger uncertainty $\sigma^{(d)}$ will penalize the corresponding weight $w^{(d)}$, and higher regularization weight λ strengthens the penalization.

Figure 2 illustrates the suppression and shifting of attribution due to different $\sigma^{(d)}$ and λ values³. In this example, since $\sigma^{(1)} > \sigma^{(3)} > 0$, $w^{(1)}$ is suppressed more than $w^{(3)}$. The total amount of attributions is conserved during the process to maintain the explanation output faithfulness to the predictor output. The suppressed explanation will be less dependent on uncertain features and less sensitive to changes in the uncertain features, so the explainer will be more robust to uncertainty. This can mitigate users' worry about uncertainty since the explained result remain similar even if a future measurement is noisily different.

Since LIME is a model-agnostic explainer that only requires input and output data from the predictor to generate explanation, modifying the LIME explainer will not affect the predictor. The benefit of this is that users can suppress attribution uncertainty only with the access to querying the predictor inference instead of modifying or retraining the predictor. However, an inherent weakness of LIME is that it is a post-hoc estimation of the model, and regularizing post-hoc explanation does not change the underlying predictor behavior, thus the predictor may still be sensitive to input uncertainty. There could be a discrepancy between the uncertainty-aware Regularized Explainer and the underlying predictor. Next, to address this issue, we propose to handle this discrepancy by regularizing predictor directly.

³The selection of λ can be personalized to user preference. It can also be selected by expected faithfulness distance (see Section 4.3 and Appendix B)

3.4. Suppressing Attribution Uncertainty in the Predictor

Several methods have been proposed to improve the human-interpretability of models by adding regularization constraints on the predictors during training. Regularization terms include elicited user preference for specific feature attributions [64], attribution priors such as smoothness in image saliency maps [22], the number of rules and rule length in a rule list [51], and depth of a decision tree explanation [75]. Here, we regularize the predictor to suppress attribution uncertainty. Regularizing the predictor by the attribution uncertainty directly suppresses the predictor model’s dependency on uncertain input features. The explanation generated from this regularized predictor should have a smaller discrepancy and higher faithfulness, and yet be uncertainty-aware. However, this requires access to the training process of the predictor; this may not be possible if the prediction and explanation are not implemented by the same developer or stakeholder.

Similarly to how we regularized the explainer with an uncertainty penalty, we define the loss function to train the Regularized Predictor $\tilde{f}(\cdot)$ with a penalty based on input uncertainty as

$$\sum_{\langle x, y, \sigma \rangle \in \mathcal{D}} \mathcal{L} \left(\tilde{f}_\theta(x), y \right) + \underbrace{\lambda \|\sigma \circ \omega(x, \tilde{f}_\theta)\|_2^2}_{\text{Attribution Uncertainty penalty}}, \quad (4)$$

where (x, y) is the input features and output label of a training instance from training set \mathcal{D} , and σ is the input uncertainty of x , $\mathcal{L}(\cdot)$ is the loss between predictor and ground-truth label, $\omega(x, \tilde{f}_\theta)$ is the attribution explanation function which takes in the predictor model \tilde{f}_θ and query instance x , and outputs the attribution $\omega^{(d)}$ for each feature $x^{(d)}$. Note that while Eq 3 is expressed in terms of linear weights from a LIME linear model explainer, we generalize the attribution explanation technique in the regularization term of Eq 4. However, note that the choice of explanation technique is limited to the type of model and its optimization methods.

Further note that, for optimal training, $\omega(x, \tilde{f}_\theta)$ should be a function of the model parameters θ . This allows the use of many search optimization methods, such as gradient descent which calculates the derivative of the loss with respect to θ . In this case, the LIME explanation is not suitable, because the linear weight in LIME is model-agnostic and not expressible in θ . Moreover, it is expensive to use LIME to compute attribution explanations,

since it needs to sample and train a linear explainer model for every training instance when the loss is computed during the training of the predictor model \tilde{f}_θ . To avoid these issues, we used a fully-connected neural network model and explain it with Integrated Gradients [72], a common technique used to explain neural networks. Integrated Gradients integrates the model output gradient over each input dimension to get the attribution scores of input dimensions. It is fast to calculate during model training and suitable for gradient descent optimizers because it is differentiable with respect to the predictor model parameters θ . We leave the generalization to other types of models or attribution explanation techniques for future work. We can use the same sampling method to estimate the attribution uncertainty generated by Integrated Gradients and show it by violin plots.

Having defined techniques to manage uncertainty, we next characterize and evaluate their performance in two experiments — a simulation study to evaluate the Regularized Uncertainty explanation methods and a user study with human subjects to compare the impact of Showing and Suppressing uncertainty on trust and decision making.

4. Characterization Study of Regularized Explanation

We seek to understand how Regularized Uncertainty explanations can suppress uncertainty in attribution explanations and whether they are more faithful under input uncertainty than baseline explanations. Thus, we conducted a simulation study with a real-world dataset, where we simulated a large number of hypothetical instances to represent different input uncertainty from query instances.

4.1. Dataset and Modeling Task

To align the simulation study with the subsequent user study, we used a real-world dataset, the UCI wine quality dataset [14], for our simulation study. This dataset characterizes 1599 red wines by their quality score (label) based on 11 chemical properties (input features). We split the dataset into a training set (80%) and a test set (20%). We focused on a subset of 5 features (alcohol, pH, total SO₂, sulphates and volatile acidity) because we want to simplify the task for users in the later user study, and the simulation study should have consistent settings. Instead of extensive testing more datasets whose input uncertainty is unknown or uncontrollable, we chose to further evaluate user perception with a human-subjects study later.

Instead of testing with only one uncertain feature, we selected two features — alcohol and volatile acidity — to be made uncertain to align with requirements for the later user study; if only one feature is uncertain, it would be too simple and predictable for participants, whereas too many uncertain features would lead to high cognitive load and user confusion. We perturbed the features for two levels of input uncertainty — high (at 1 standard deviation), medium (at 0.5 standard deviations) and low (at 0.3 standard deviations). For each instance in the dataset and each uncertainty level, we synthesized 50 hypothetical data points by perturbing the two features.

Regarding the modeling task, although most interpretable machine learning models are studied with classification problems, we evaluate with regression models, so that we can precisely measure the similarity and differences in the explainer’s, predictor’s and user’s inferences. To generalize regression to binary classification, a logistic function can be applied.

4.2. Trained Models

Depending on the explanation regularization approach, we trained various predictor and explainer models: a 3-layer neural network model as the baseline predictor ($f = \text{NN}$), baseline and regularized LIME explainers on the baseline predictor ($g_f = \text{LIME}(\text{NN})$, $\tilde{g}_f = \text{RegLIME}(\text{NN})$); the regularized predictor ($\tilde{f} = \text{RegNN}$) and Integrated Gradients explanation on the baseline and regularized predictors ($g_f = \text{IG}(\text{NN})$, $g_{\tilde{f}} = \text{IG}(\text{RegNN})$). Note that the predictor model of $\text{IG}(\text{RegNN})$ is both regularized and explained by Integrated Gradients. Since two different explanation techniques, LIME and IG, were used to regularize the explainer $\text{RegLIME}(\text{NN})$ and predictor $\text{IG}(\text{RegNN})$, in order to control the study, the two regularized methods are compared with their baselines LIME(NN) and IG(NN), respectively. Next, we evaluated the performance of the different explanations by defining an Expected Faithfulness metric.

4.3. New Measure: Expected Faithfulness

To investigate explainer faithfulness under uncertain inputs, we have to measure how the explainer can make predictions that agree with the predictor even if the instance has uncertainty or noisy measurements. We defined a

new measure — *Expected* Faithfulness Distance^{4,5} — to measure this effect and show how expected faithfulness distance can be reduced with Regularized Uncertainty explanation to improve expected faithfulness.

The inference of an explainer should be consistent and similar to the inference of the predictor; i.e., explainers should be faithful to predictors [49, 57, 63]. For example, the LIME explainer is a linear model trained to infer the same output as the underlying predictor. The faithfulness distance of baseline explainer for an instance x is defined as the difference between the predictor’s inference $f(x)$ and the baseline explainer’s inference $g_f(x)$, i.e., $F_0 = (f(x) - g_f(x))^2, F_0 \geq 0$. This is a point-estimate metric. When the input feature of instance x is uncertain, we specify that a locally robust explainer should be less reliant on input noise ϵ , thus the explainer’s inference of the noisy input $x + \epsilon$ should be the same as the predictor’s inference on the original instance x . To measure this local robustness, for the baseline explainer g_f , we define the faithfulness distance under uncertainty as $F_{g_f} = (f(x) - g_f(x + \epsilon))^2$.

For a baseline explanation technique i.e. LIME(NN) or IG(NN), we expect that on average the baseline is less faithful at explaining uncertain instances because its explanation is sensitive to noise. We define its average faithfulness distance over uncertain inputs as the *expected faithfulness distance*:

$$E[F_{g_f}] = E_\epsilon [(f(x) - g_f(x + \epsilon))^2] = \int_{-\infty}^{\infty} (f(x) - g_f(x + \epsilon))^2 P(\epsilon) d\epsilon, \quad (5)$$

and assert that $E[F_{g_f}] \geq F_0$. In the simulations, we generated 150 hypothetical instances around each test instance by sampling ϵ from a Gaussian distribution at high (standard deviation = 1), medium (standard deviation = 0.5) and low (standard deviation = 0.3) uncertainties. See the proof in Appendix A.

On the other hand, for the Regularized Uncertainty explainer $\tilde{g}_{f(\cdot)} =$, *RegLIME(NN)*, we expect that its faithfulness to explain uncertain instances

⁴Note that higher distance actually refers to less similarity, and lower faithfulness, so this should technically be called ”Unfaithfulness Distance”, but we term it as ”Faithfulness Distance” to simplify nomenclature.

⁵We call this metric a “distance” instead of “difference”, since we calculate it as the aggregate mean squared error (MSE) of differences for all instances, to represent global faithfulness for the explainer model.

is better than the baseline explainer's point-estimated faithfulness distance F_0 , due to the Regularized Explainer's robustness. First, we specify that the robust inference outcome should aim to be similar to the prediction for the original instance, i.e., $\tilde{g}_f(x + \epsilon) \approx f(x)$. The faithfulness distance for the Regularized Explainer on an uncertain input instance is thus $F_{\tilde{g}_f} = (f(x) - \tilde{g}_f(x + \epsilon))^2$. Although the explanation at some uncertain instances may be less faithful than at the original query instance x , on average across all hypothetical uncertain instances, the faithfulness distance of $\tilde{g}_f(\cdot)$ may sometimes be as good as or better than that of the point-estimated explanation on original query instance x , i.e., $\text{Prob}(E[F_{\tilde{g}_f}] < F_0) > 0$. For the Regularized Explainer, the expected faithfulness distance is:

$$E[F_{\tilde{g}_f}] = E_\epsilon [(f(x) - \tilde{g}_f(x + \epsilon))^2] = \int_{-\infty}^{\infty} (f(x) - \tilde{g}_f(x + \epsilon))^2 P(\epsilon) d\epsilon \quad (6)$$

For the Regularized Predictor $\tilde{f}(\cdot)$ we defined, its prediction on a query instance x is explained by its model-dependent explanation technique $g_{\tilde{f}}(\cdot)$, and $g_{\tilde{f}}(x)$ is the explanation inference of instance x inferred by the explanation. Here the outcome is calculated by summing up all attributions and model bias. The point-estimate explanation faithfulness distance F_0 is defined as before, i.e., $F_0 = (f(x) - g_f(x))^2$. Note that here the baseline explanation is IG(NN). Thus, we define the expected faithfulness distance over uncertain input for the Regularized Predictor IG(*RegNN*) as:

$$E[F_{g_{\tilde{f}}}] = E_\epsilon [(f(x) - g_{\tilde{f}}(x + \epsilon))^2] = \int_{-\infty}^{\infty} (f(x) - g_{\tilde{f}}(x + \epsilon))^2 P(\epsilon) d\epsilon \quad (7)$$

Similar to the Regularized Explainer, we hypothesize that $E[F_{g_{\tilde{f}}}] \geq F_0$ for the baseline predictor $g_f = \text{IG}(\text{NN})$, and $\text{Prob}(E[F_{g_{\tilde{f}}}] < F_0) > 0$ for Regularized Predictor $g_{\tilde{f}} = \text{IG}(\text{RegNN})$.

4.4. Results: Regularized Explainer and Regularized Predictor Improve Expected Faithfulness Distance

Next, we conduct a simulation experiment to investigate how likely the explanations of instances are to benefit from improved expected faithfulness distance by suppressing uncertainty with the Regularized Explainer and Regularized Predictor.

Figure 3 shows the probability of $E[F_{\tilde{g}_f}] < F_0$ being true, which indicates how often the Regularized Explainer *RegLIME*(NN) is more faithful than the

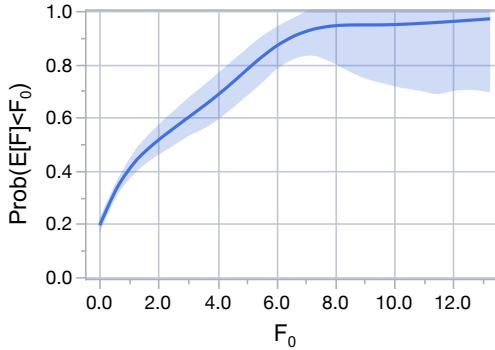


Figure 3: Results of simulation on wine dataset. X-axis F_0 is the point estimate faithfulness distance from Baseline explainer LIME(NN), which is the distance between explainer prediction and model prediction. Y-axis is the probability of Regularized Explainer *RegLIME(NN)* has a smaller expected faithfulness distance on all hypothetical instances than the point estimate faithfulness distance of Baseline (F_0). The shaded area is the standard error. This result is calculated by the high input uncertainty level (Gaussian distribution with 0 mean and 1 standard deviation).

Baseline explainer LIME(NN) under input uncertainty. The y-axis value, calculated by the number of instances, satisfies $E[F_{\tilde{g}_f}] < F_0$ divided by the number of instances in the dataset. And x-axis represents the different values of baseline explainer faithfulness distance F_0 for different x_0 . We can see that for less faithful baseline explainers (larger F_0), regularizing the explainer is more likely to have smaller expected faithfulness distance under uncertainty (i.e., $Prob(E[F_{\tilde{g}_f}] < F_0)$ gets higher). This could be because when F_0 is small, the baseline explanation is already very faithful, so there is not much room for regularization to improve.

Similar to Figure 3, Figure 4 shows that regularizing on the predictor helps improve the Expected Faithfulness (higher $Prob(E[F] < F_0)$), especially for predictors with poorer baseline explanations (higher F_0). Note that the range of F_0 for explainers of the Regularized Predictor IG(*RegNN*) is much smaller than that for Regularized Explainer *RegLIME(NN)*, which means that the baseline Integrated Gradient is a more faithful explanation method than baseline LIME. Although the Regularized Predictor has better results on expected faithfulness than Regularized Explainer, regularizing the predictor at training is more computationally expensive than regularizing the explainer at inference time, and in many scenarios, developers may not have access to change the predictor. In contrast, regularizing the model-agnostic explainer

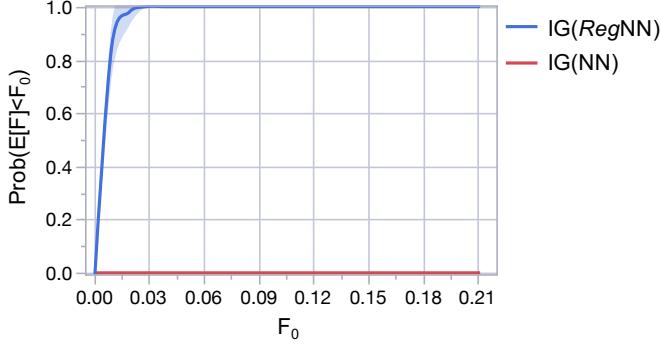


Figure 4: Results of simulation on Regularized Predictor. X-axis F_0 is the point estimate faithfulness from Baseline predictor $\text{IG}(\text{NN})$. Y-axis is the probability that an explanation has a smaller expected faithfulness over all hypothetical instances than the point estimate faithfulness of explanation from the Baseline predictor. Blue is for the Regularized Predictor $\text{IG}(\text{RegNN})$, $\text{Prob}(E[F_{g_f}] < F_0)$. Red is for IG explanation on baseline neural networks predictor $\text{IG}(\text{NN})$, $\text{Prob}(E[F_{g_f}] < F_0)$. Error bar is standard error. This result is calculated by the high input uncertainty level (Gaussian distribution with 0 mean and 1 standard deviation).

enables to suppress attribution uncertainty based on users' preference for different uncertainty handling strategies.

4.5. Results: Regularized Explainer and Regularized Predictor Have Similar Explanations

We have shown that both Regularized Explainer $\text{RegLIME}(\text{NN})$ and Regularized Predictor $\text{IG}(\text{RegNN})$ methods can generate explanations that are robust to uncertain input and more faithful than their baselines $\text{LIME}(\text{NN})$ and $\text{IG}(\text{NN})$, but how similar or different are the explanations from these methods?

Figure 5 shows four variants of explanations for an instance example. While predictor is explained by different attribution explanation techniques i.e., LIME and IG regularizing either on the explainer or predictor produce a similar suppressed explanation result. The Regularized Explainer and Regularized Predictor methods can suppress attribution to uncertain features in a similar way resulting in similar attribution explanations. To quantify this similarity across all instances, we calculate the similarity between explanations based on the Euclidean distance of their feature attributions.

Figure 6 shows how explanations from the Regularized Explainer $\text{RegLIME}(\text{NN})$

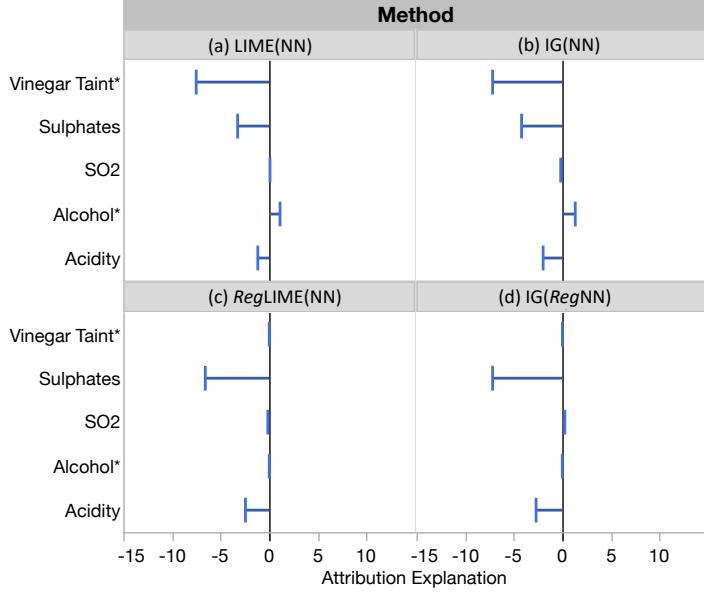


Figure 5: Example explanations from various techniques: (a) Baseline LIME explainer on baseline predictor LIME(NN); (b) Baseline Integrated Gradients explanation on baseline predictor IG(NN); (c) Regularized LIME explainer on baseline predictor *RegLIME*(NN); (d) Integrated Gradients Explanation on Regularized Predictor *IG(RegNN)*. The two baseline explanations are slightly different, but the two regularized explanations are similar after suppression. * indicates that the Vinegar Taint and Alcohol features had input uncertainty.

and Regularized Predictor *IG(RegNN)* are also similar. Their similarity is as close the similarity between explanations from the baseline Explainer LIME(NN) and baseline Predictor IG(NN) (compare Figure 6 left two bars), and closer than the similarity between baseline and regularized explanations by more than 5 times⁶ (contrast Figure 6 first left bar with right two bars). The differences between regularized and baseline explanation distances were significant in Wilcoxon Rank Sum tests ($p < .0001$) with large effect sizes ($r = .569$ to $.579$) [62].

A take-away from this analysis is that uncertainty-aware explanations from the model-agnostic Regularized Explanation can be as faithful as those from Regularized Prediction. The former method also has the benefit of being easier to train.

⁶ $3.28/0.599=5.48$, $3.24/0.599=5.41$, $3.28/0.648=5.06$, and $3.24/0.648=5$, respectively.

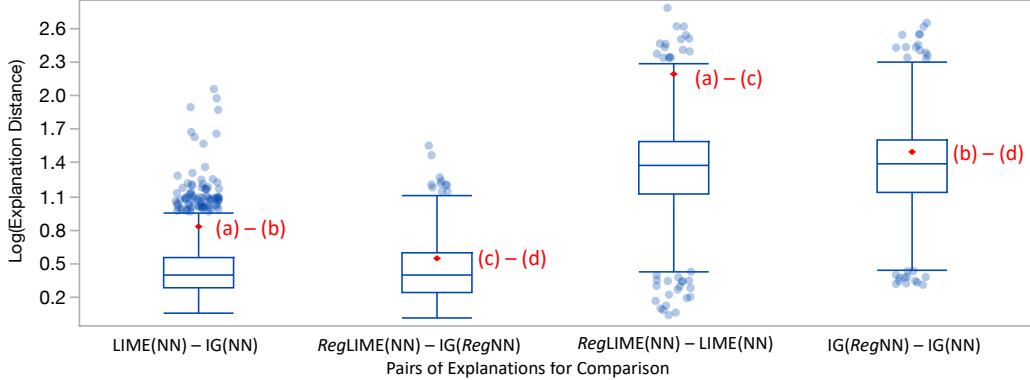


Figure 6: Distribution of logarithm of explanation distance between different techniques, shown as box plots (1st quartile, median, 3rd quartile) and whiskers (1st quartile - $1.5 \times$ interquartile range or min, 3rd quartile + $1.5 \times$ interquartile range or max). The two baseline explanations are as similar to each other as the regularized explanations to each other (low distance in two left plots). The suppressing effects on attributions by regularizing the explainer and regularizing the predictor are similar too (similar distances in two right plots). Red diamond markers indicate explanation distances for pairs of explanations shown in Figure 5. For example, (a)-(b) represents the logarithm of Euclidean distance between the explanations in Figure 5(a) and Figure 5(b). We used logarithmic transform to provide a compact visualization of the heavily skewed distribution of results.

5. User Studies to Show and Suppress Uncertainty in Attribution Explanations

We conducted a formative qualitative study and follow-up summative quantitative study to evaluate the two proposed explanation techniques to handle attribution uncertainty. With the qualitative study, we sought to understand whether and how users could benefit from using either Show or Suppress, how different approaches affect users’ trust in the AI and decision making, and how users interpret and use the explanations. With the controlled quantitative user study, we compared the Show and Suppress explanation techniques against baselines to evaluate at scale how using them influences user trust, perceived helpfulness and decision quality on a rating and decision task.

For consistency, we employed the same experiment task, experiment apparatus and experiment variables for the qualitative and quantitative studies. That is, the experiment task was used as a probe in the qualitative study.

In this section, we first describe the user task, independent variables, control variables, system and explanation user interface (UI) apparatus. We will describe the method, procedure, and results for the qualitative and quantitative studies in the subsequent sections.

5.1. User Study Task: Wine Quality Control Inspection

We designed a user decision task that involves estimating a score to make a binary yes/no decision. We defined a scenario to fit the context of the UCI Wine Quality dataset that correlates chemical properties with wine score ratings labeled by human experts. We presented participants with the scenario that they work as a quality control inspector to estimate the quality score (0 to 100) of wine at a production winery and decide whether to accept or reject a wine for sale. If the participant thought the wine score is >50 , then he should accept it, otherwise, he should reject (i.e., reject ≤ 50). Participants were introduced to the Smart Wine Rating System that estimates (predicts) the wine quality score from 0 to 100 based on five chemical properties, and told that they could use the system to help with decision making. Participants are taught about how to read the system user interface (described in Section 5.3) and also informed about the uncertainty in chemical readings. Depending on the experiment condition, the System provides basic information (readings with uncertainty and predicted quality score with uncertainty), or various forms of additional information. We did not explicitly call these “explanations” to the participants. Participants were told that the wines they were inspecting are borderline cases⁷ and although the system predicts a score, it may be wrong. Wine instances were chosen such that 50% of them should be rejected. For each decision task, participants indicate their inspection decision, estimated wine score value and uncertainty of their score as a number.

More details of the scenario as presented to participants are described in C.3. In the qualitative study, we emphasized on understanding the participants’ thought processes and preferences across many explanation techniques,

⁷Cases were chosen to emulate the paradigm where the machine learning system would defer to human judgement if it is not confident of its prediction. Hence, with uncertainty, most cases (29/30) had confidence intervals straddling the mean value, i.e., lower bound below mean and upper bound above mean. We expect that participants would make quick decisions that would likely agree with the prediction for non-straddling cases and perceive them to be easy.

so they were asked to take their time to study the UI and think aloud while doing so, and participants were exposed to three system variants. For the quantitative study, we emphasized quick and intuitive decision making, and applied a timed incentive to make at least 12 correct decisions for 15 wines within 15 minutes for each system variant. Participants used two variants of the system one after the other, and were told to aim for at least 80% correct decisions (i.e., $\geq 12/15$) in their inspections each time.

We chose this task for several reasons. First, we wanted to minimize the influence of prior knowledge and personal preferences in familiar tasks. With this task, lay people are not familiar with rating wines based on their chemical properties, and we can control the knowledge by training. Second, it is plausible that inputs can be uncertain due to errors when measuring chemical properties. Third, predicting wine quality rating is a classic regression problem in machine learning, and this allows us to more precisely measure the impact on user labeling at a higher resolution than only on a classification task. The dataset and model settings are consistent with the Simulation experiment.

5.2. Experiment Treatment: Independent, Control, and Dependent Variables

The experiments focus on one primary independent variable about which explanation technique is used for decision making:

1. Technique (5 levels): score prediction system with no explanation (None⁸), attribution explanation with subscore values in a tornado plot (Baseline, see Figure 7a)), tornado plot including subscore uncertainties (Show, see Figure 7b), tornado plot of suppressed subscores (Suppress, see Figure 7c), tornado plot of Suppressed subscores and shown suppressed uncertainties (ShowSuppress⁹, see Figure 7d).

We treat Technique as a within-subjects design, thus we have 5 conditions in the experiment. Each participant uses more than one technique with

⁸We included the None baseline as a reference to verify whether any effects are due to explanations in general or about the specific explanation technique, but we focus our analysis on the four explanation techniques.

⁹We only used the Regularized Explainer to control the behavior of the predictor model and keep it consistent across explanation conditions; regularizing the predictor will lead to different predictions for the same instances and this is a confounder that affects the user's trust [77]. Nevertheless, our simulation results have shown that the Regularized Explainer has similar explanations as the Regularized Predictor.

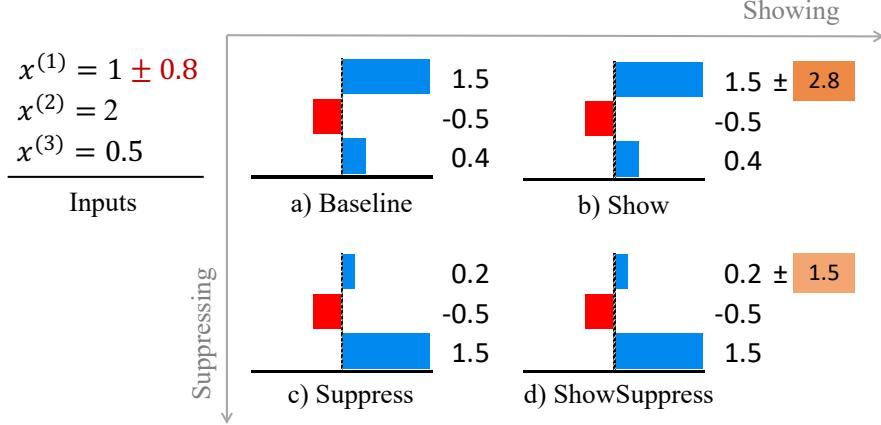


Figure 7: Conceptual examples of simplified visualization in user study.

multiple trials per technique. To limit experiment variability and present challenging cases for user intervention to simulate when an AI system defers decisions to humans due to its lack of prediction confidence, we set the following variables constant as described.

2. Input Features with Uncertainty (Alcohol and Vinegar Taint): we set only two input features to have uncertainty, and with equal relative amounts. All other input features have no uncertainty.
3. Input Uncertainty (medium): we fixed the uncertainty of each input feature as a Gaussian distribution with a standard deviation that is half the standard deviation of the feature values.
4. System Correctness (all instances' system prediction is correct): For each selected instance, we made sure the predicted and actual (ground truth) scores to be consistent, i.e. on the same side of the decision threshold (i.e., both above or both below).
5. Closeness to Decision Threshold (between 40 and 60, i.e., $| \text{system score} - 50 | < 10$): For all selected instances, we made sure the system prediction score is close to the decision threshold 50 (within 40 to 60) to present challenging ambiguous cases.

To choose a representative sample of hypothetical instances with uncertainty, we processed and selected data based on the following steps:

- i. For each instance from the dataset in the simulation study, add noise according to a Gaussian distribution to create 50 hypothetical instances

- with uncertainty.
- ii. Filter instances with the aforementioned control criteria.
 - iii. Divide instances into separate clusters by performing hierarchical clustering on their feature values, attribution explanations, model predictions, the uncertainty of these values and ground truth scores.
 - iv. Perform stratified sampling across clusters to select final instances for the experiment.

Ultimately, we selected 34 wine instances for our user studies, 4 for practice, and 30 for the main trials. The mean Closeness to Decision Threshold is 1.95 (SD=1.05).

We measured user performance *per trial* and overall system impression of the explanation technique *per condition*. For each trial, we measured

1. Decision Time: the amount of time in seconds for the participant to decide to accept or reject the wine, estimate the wine quality score and score uncertainty. This includes time to see and study the input values and uncertainty, system prediction and explanation (if available). Decision Time is only recorded and analyzed for the quantitative study.
2. User quality score value: the wine quality score estimated by the participant, denoted as $Score_{User}$.
3. User quality score uncertainty: the \pm error bound of the participant's wine quality score value, denoted as $ScoreUncertainty_{User}$. This should refer to half the 90% confidence interval width to correspond to what the system displayed.

For each condition, we measured

4. Confidence: the participant's self-reported confidence in decision making on a 7-point Likert scale from strongly disagree to strongly agree.
5. Trust: self-reported trust of the system for an accurate score prediction, on a 7-point Likert scale from strongly disagree to strongly agree. Note that the participant may trust the system, and yet not feel confident about her decision due to other factors.
6. Helpfulness: self-reported overall helpfulness of the system and any available information or explanations for decision making, on a 7-point Likert scale from strongly disagree to strongly agree.
7. Helpfulness of specific system features: self-reported helpfulness of different features of the system on 7-point Likert scales from strongly disagree to strongly agree. System features include reading value, read-

ing uncertainty, subscore value, subscore uncertainty, suppressed subscore value, suppressed subscore uncertainty, score value, and score uncertainty. These helpfulness questions were asked based on Technique condition.

These measures were recorded for the qualitative interview and quantitative study and used to compute dependent variables (See Section 5.6.2).

5.3. Experiment Apparatus

We designed a simplified user interface to communicate attribution explanations with uncertainty for the Show and Suppress techniques to be usable by lay users from Amazon Mechanical Turk. We prioritized usability and familiarity over more precise and expressive visualizations. In an on-line pilot study with the visualization in Figure 1, we found that many lay users had difficulty in learning to read the violin plots to pass comprehension tests. Hence, we chose a familiar table layout with input values as numbers, attribution values, and uncertainties represented as (\pm) numbers. Uncertainty numbers indicate the 90% confidence interval half-width, suggesting the range within which the true value has a 90% chance to lie. with small violin plots. To show the attribution explanation, we kept the tornado plot, but as a simple bar chart embedded in the table, rather than a large T-bar chart. Attribution uncertainty is simply presented as a single number, rather than a distribution curve or error bars. For simplicity to lay users, we call the attribution values as “subscores”.

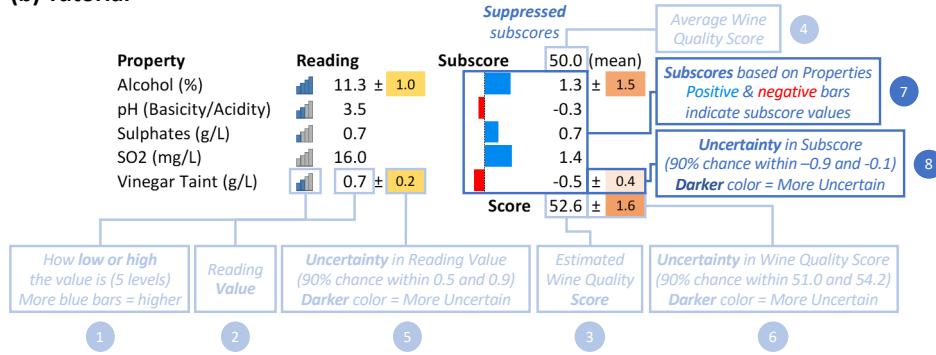
Figure 8 shows the modular table user interface that can show different UI components depending on the explanation technique condition. 1) & 2) the chemical property reading value, which is the system input, and the indicator of how high or low the reading value is; 3) the estimated wine quality score, which is the system output; 4) the average wine quality score for all the wine, which is also the decision boundary; 5) the uncertainty of the property reading values, which is the input uncertainty; 6) the uncertainty of wine quality score, which is the output uncertainty; 7) subscores for each property, which is the attribution explanation; and 8) the uncertainty in subscores, which is the attribution uncertainty. All participants only learned the system features 1) to 6) in the training, and would not learn about the system explanation until the main study.

The experiment was implemented as an online Qualtrics survey with different images for each system UI trial, and different trials on subsequent pages. We describe the experiment procedure next.

(a) Wine Knowledge

Chemical Property	Generally, wines with ...	Reading	
		Min	Max
Alcohol (%)	Higher % alcohol has higher quality . But too much (>14) alcohol reduces quality.	8.3	14.9
pH (Basicity/Acidity)	Higher pH (less acidic) has slightly lower quality .	2.74	4.01
Sulphates (g/L)	More sulphates has higher quality , by controlling tartness and clarity. But too much (>0.85) slightly reduces quality.	0.33	2.00
SO ₂ Sulphites (mg/L)	More preservative sulphites has lower quality .	6	289
Vinegar Taint (g/L)	Higher vinegar taint has lower quality .	0.12	1.59

(b) Tutorial



(c) Main Trial Interface

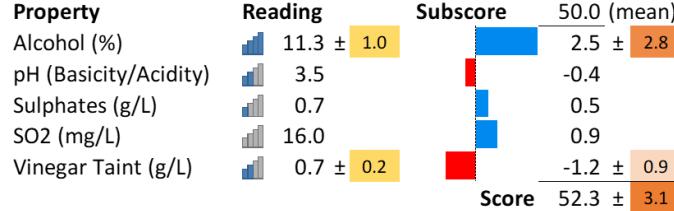


Figure 8: (a) Background Knowledge table shown to participants to teach them how each chemical property affects the wine quality score. The description is obtained by analyzing partial dependence between input features and the rating score, where a trend is flat if the slope of between wine quality score and the standardized property is less than 2.0/unit. (b) Annotated screenshot of the system user interface (ShowSuppress version with all 8 components displayed) used to teach participants on how to read the UI properly. Other system variants will have fewer UI components. (c) System UI as seen during practice or main trial without annotation clutter (Show version displayed). With the Suppress and ShowSuppress conditions, participants can hover the mouse over the figure to compare the explanation before and after suppression (subscore and score values and uncertainties will switch). Yellow and orange highlights indicate the amount of uncertainty in the reading and subscore, respectively.

5.4. Experiment Procedure

We describe the procedure for participants in the quantitative study, then describe the truncated procedure for qualitative study participants. In the quantitative study, participants from Amazon Mechanical Turk follow the procedure:

1. Read a welcome message and consent to the study.
2. Study a tutorial and answer corresponding screening questions on:
 - i. The task description and task background knowledge (Figure 8a).
 - ii. System user interface versions with the incremental introduction to input readings and output score, readings and score uncertainty (Figure 8b).
3. If participants had <5 out of 7 screening questions correct, they will be disqualified from the study. Those that pass proceed to the next step and answer 6 questions about Uncertainty Tolerance and Uncertainty Decisive (7-point Likert Scale, see Figure C.7).
4. Task reminder to describe the user task and introduce bonus incentives (12 correct at \$0.25, with \$0.25 increments until 15 correct at \$1.00).
5. Randomly assigned to two (of 5) explanation Technique conditions, where for each condition, the participant sees:
 - i. Pre-Condition tutorial on the specific explanation UI (Figure 8b, Figure C.6), with comprehension test questions for data quality checking, not to screen participants.
 - ii. Two practice trials, where for each trial, the participant
 - Estimates his score value and uncertainty¹⁰ using sliders¹¹, and
 - Decides to accept or reject the wine.
 - Sees a displayed timer to indicate how much time has passed.

¹⁰We asked for a simple number, and did not ask participants to estimate a 90% confidence interval to avoid technical jargon and they have already learned from the tutorial that the uncertainty refers to a 90% chance of the value being within the range. We had piloted asking uncertainty with the bin and balls question [28], but wanted a much faster data entry method, and did not need very high accuracy.

¹¹We recorded with sliders instead of number text entry to speed up data entry and allow easy entry of mixed decimal numbers (see Figure C.9).

- Sees a second page (See Figure C.10) that reviews her answers, gets feedback on the ground truth wine score and whether her decision of accepting/reject is correct, and answers a question about her rationale¹² in the decision and rating estimation. This page was not included for the main trial to not interrupt rapid task completion.
- iii. Task and incentive bonus reminder.
- iv. 15 main trials (randomly ordered) with the same format as practice trials but without the second page of feedback.
- v. Post-Condition page where the participant reviews her performance (how many trials correct) and answers questions to rate her Confidence in decision making, Trust in the system estimation, and Helpfulness of the system information for decision making.
- vi. The participant takes a one-minute break after finishing the first condition.
6. After both conditions, the participant answers demographics questions (e.g. age, gender, ethnicity, education, employment status and occupation) and receives feedback on how much bonus she would get.

The qualitative study is conducted online via a Zoom audio call with screen capture recording, which the participant consents to. They also go through the tutorial and can ask the experimenter clarification questions. They do not receive a bonus for their task performance, since we do not incentivize for speed and accuracy. Finally, the focus of inquiry was to observe their usage and rationale using the think aloud-protocol [61] and interview questions, rather than their answers to the rating questions. Mummolo et al. found little evidence for the experimenter demand effect for survey studies on Amazon Mechanical Turk [60]. Nevertheless, to control for the demand effect, we randomized the order of testing explanation techniques, regularly informed participants at every trial that the system could be wrong, so that they would not necessarily want to copy or trust the system.

¹²Her answers are repeated for reference. The rationale question serves two purposes: i) manipulation check to ensure that users are paying attention and reasonably understand the user interface, ii) stimulate users to self-explain to raise their level of understanding before proceeding to the main trials.

5.5. Qualitative Study: Usage and Usefulness of Showing and Suppressing Attribution Uncertainty

Although many prior studies have investigated the benefits and limitations of showing explanations of AI, it is unclear how communicating uncertainty in explanations impacts the user experience. In particular, we propose two different explanation techniques to handle attribution uncertainty, so it is important to understand how users will comparatively perceive and understand, and apply them. Therefore, we conducted a qualitative study by observing how users made decisions on the wine inspection task and interviewing their experience of different explanation techniques.

In particular, we were interested in the following objectives:

1. Verify that users can easily acquire a basic understanding of the baseline tornado plot for attribution explanations. We also explored any further needs on the explanation of the AI.
2. Pilot a training method to teach users how to understand the different explanation techniques (Show or Suppress) and identify usability issues and misunderstandings of the visualizations. This helped us to further improve the experiment apparatus for evaluation with remote MTurk workers.
3. Observe and enquire how users interpret and employ the various information (UI component) in each explanation technique to estimate their wine score and uncertainty, and to make decisions. For example,
 - Showing: Would showing subscore uncertainty influence the user's score estimate, would she use both sides of the uncertainty bounds? Would showing uncertainty raise or lower the user's confidence?
 - Suppressing: Would suppressing subscores and subscore uncertainties raise the user's confidence and trust? How would it impact user's score estimation? For each objective, we ask follow-up questions to gain insight into the interviewees' rationale.
4. Whether and why users have a preference for different explanation techniques, and for what circumstances.

We recruited 12 interviewees from a university mailing list, aged 21 to 53 ($M=25.9$), 9 females. 6 were undergraduate students, 5 were graduate students, and 1 was alumni. They studied various majors, including nursing, medicine, psychology, real estate, biology, accountancy, chemical engineering, mechanical engineering, project and facility management, and computer sci-

ence. Interviewees were conducted via the Zoom call with a shared screen to maintain social distancing during the COVID pandemic. Audio and screen movements were recorded with participant consent. Interviews took 35 to 57 minutes ($M=48.5$) and participants were compensated with a \$10 SGD Starbucks gift card for their time.

Through a semi-structured interview with participants using the explanation techniques in a web interface, we conducted the qualitative study with the following procedure: After consenting to the study, participants were introduced to the wine inspection task, and went through the relevant tutorials as described in Section 5.3. Each participant rated 1 to 2 wines, and, for each wine, sequentially used three out of four explanation techniques with different orders (e.g., Baseline → Show → ShowSuppress, Baseline → Suppress → ShowSuppress, Baseline → Show → Suppress, Baseline → Suppress → Show). We assigned interviewees to different arrangements based on whether we have saturated on our understanding of comparison between different explanation techniques. We asked the participant to think aloud while using each explanation technique. After each trial (one technique for one instance), we asked users questions regarding our aforementioned objectives. Due to limited time, we only asked questions that were relevant to the interviewee’s comments and that we had not saturated on. We ended by collecting demographic and background information.

Next, we discuss our findings and organize them in several themes.

5.5.1. Understanding of Baseline, Show & Suppress Attribution Explanations

We probed the participants’ mental models of how the AI system works. We found that most participants understood that the system makes predictions by adding up the subscores in the attribution explanation of all chemical properties to the total quality score, and that subscores can contribute positively or negatively. This is consistent with our tutorial. P6 understood that “*it gives some weightage for each 5 components, and the weightage for sulphites seems to be quite large (subscore = -2.9) ... It shows quite clearly how those properties contribute negatively or positively to the overall wine quality.*” Participants could comfortably describe their decision-making process using the system subscores; e.g., P10: “*For alcohol, I can see the subscore is highly negative (-6.1). Then for pH (0.3), Sulphates (0.9) and SO2 (1.1) are all slightly positive. But Vinegar Taint (-1.3) is slightly negative. ... Based on these, I will likely reject this wine, because the contribution of Alcohol and Vinegar Taint in the negative range plays a very high contribution to the total*

score."

P10 found that seeing subscores and tornado plot increased her understanding of the system "because the reading and range of each property are quite different, we are not sure whether the value contribution is high or low to the final score and how much the reading contributes to the final score, but with the subscore we can get how much each property contributes to the final score. And by comparing the height of the red bars and the blue bars, it can allow us to make the final decision especially when it comes to uncertainty." From the interview, we learned that most participants are able to understand the model by reading attribution explanations, used the attribution to make decisions, trusted or questioned the system because of the attribution explanation.

However, when relating the system subscores to their general knowledge from the background table (Figure 8a), some participants questioned the validity of subscore values in the system. From the background table, P12 learned how the chemical property reading values could affect the wine quality score, and disagreed with the system on one of the subscores: "*The percentage of the Alcohol is in the lower range, so I think it will affect more (than other chemical properties). . . . I think the alcohol subscore should be affecting it even bigger like -2.8 (instead of the system shown -2.1).*"

Participants generally understood the Show and Suppress techniques regarding subscore uncertainty. Participants could understand how input uncertainties lead to subscore uncertainties. P6 observed that "*(the subscore uncertainty) is proportional to which property has larger uncertainty . . . there's more uncertainty associated with (the uncertain) Alcohol in determining the quality, it is helpful in that way.*" P9 noted that subscore uncertainty may not be monotonically increasing with input uncertainty, and saw "*that a small variation in alcohol can lead to larger subscore uncertainty.*" Participants understood that Suppress would reduce the subscores of uncertain input properties and reallocate the attribution. P11 understood that "*the system makes the subscores nearer to zero, so the alcohol and VT won't affect the wine quality.*", and P2 noticed that "*the Sulphate subscore (magnitude) becomes larger to compensate (the suppressing).*"

5.5.2. Showing Uncertainty Is Diversely Helpful

We found that participants had different approaches to adjust their score estimation after seeing the additional subscore uncertainty (compared to a previous baseline explanation). P8 increase his estimate of the score uncer-

tainty compared to his earlier estimate when seeing the Baseline explanation, because “*apart from (the output uncertainty being) ± 1.9 , the subscores also have high uncertainty.*” Although subscore uncertainty is actually the decomposition of the score uncertainty, and the output score uncertainty does not change, showing uncertainty gave these participants the impression that the system is more uncertain than before, thus a few participants also increased their decision uncertainty after seeing subscore uncertainty. However, perceiving more uncertainty does not mean that the participants will be less confident in their decision. Most participants did not change their confidence after seeing the subscore uncertainty, with some even increasing their confidence. P3: “*It (subscore uncertainty) makes me even more confident of my decision to reject because I have seen that alcohol (subscore) can fall anywhere within this range. It is even more uncertain, and I am more sure that I have made the safe decision to reject the wine.*” P11 increased her trust than Baseline after seeing subscore uncertainty “*because without subscore uncertainty, just by reading the reading uncertainty, I can’t really guess how the reading affects the subscore. With the subscore uncertainty, I can understand how much the subscore will change.*” She also said that with subscore uncertainty “*I am able to guess the score(output) more accurately.*”

Indeed many participants felt that showing subscore uncertainty helped them to understand the system and to make decisions. P6 mentioned that “*it seems that (the system) quantifies the uncertainty contributed by each property more clearly (than Baseline).*” Showing subscore uncertainty helped participants to make better use of input uncertainty and understand where the output uncertainty is from. P7 mentioned that knowing subscore uncertainty makes the input uncertainty more helpful than Baseline, because subscore uncertainty helps him to “*deduce*” the rating of the wine from the input uncertainty. With subscore uncertainty, P11 did “*(find) the wine score uncertainty more helpful, because I can understand better where the 1.9 (output uncertainty) comes from. Previously (in Baseline), I thought it was a random estimation of uncertainty.*”

Instead of just considering uncertainty as an error, some participants used the uncertainty to interpret one-sided bounds to consider the best-case and worst-case scenarios. P3 and P12 wanted to be safe with their accept or reject decisions, and decreased their score after seeing subscore uncertainty. According to Prospect Theory [36], they demonstrated risk aversion to weight a negative outcome as more likely. For the uncertain property readings, i.e. Alcohol, they focused on the worst subscore that Alcohol could get by

subtracting the subscore uncertainty value from their previous score estimate when viewing the Baseline explanation. On the other hand, some participants decided to accept borderline wine because the score “*has a chance to be above 50*” (P1). P1 added the subscore uncertainty to the displayed score (upper bound estimate) and increased his score estimate.

One participant, P11, used the uncertainty to provide flexibility in her interpretation to align the system behavior with her belief. For an instance with low Vinegar Taint (VT), P11 felt that VT should have a positive subscore to be consistent with the background knowledge table. However, the subscore was calculated to be slightly negative (-0.1). Having seen the VT subscore uncertainty of 1.0, she happily increased her score estimate by 1.0, since she imagined that the VT subscore could actually be positive. With subscore uncertainty, P11 was “*able to guess the score more accurately, because I think for the Vinegar Taint, if (the score) drops by 0.1, then I think it would be a bit inaccurate because (the knowledge table) says lower (VT reading value) will increase (the score), so I used the +1. I increase my score by 1 because for vinegar I plus one because I think it should increase the wine score. So I used the uncertainty of subscore to estimate my score. My estimation of VT subscore is 1.1.*” Therefore, the subscore uncertainty gave credence to her belief that the subscore could be positive, allowing her to be more assertive in her estimation.

5.5.3. Suppressing Uncertainty Generally Improves Trust and Confidence

We found that suppressing attribution uncertainty improved users’ confidence in the decision and trust of the system. Compared to Baseline, when the subscore is suppressed, P8 was more confident in her decision “*because the (score) uncertainty is lower, and I know that even if (I subtracted) 0.9 the (lower bound) score would be still above 50.*” P11 also increased her trust of the system compared to Baseline “*because I think with the suppressing, it will eliminate the uncertainty in Alcohol and Vinegar Taint measurement, it will be more accurate.*” When comparing Show to ShowSuppress versions of the system, participants also increased their confidence and trust due to the decrease in score and subscore uncertainty.

Some participants were comfortable with and appreciated the re-attribution due to uncertainty suppression. P7 thought the “*suppressed subscore is pretty good because it can readjust and recalibrate the subscores when the uncertainty is high.*” However, some participants found that this ”compensation” could be confusing because the attributions become mismatched with the back-

ground knowledge. For example, P2 compared the background table and system explanation and said: *“If I have the knowledge table, I can see that although the (Alcohol subscore) uncertainty is low, it doesn’t mean the reading is a good one. ... Because after the suppression, the subscore is –0.1. Although subscore is less now (closer to 0, higher than Baseline subscore –6.1), but the reading of the property is quite far away from the reading of good wine”*

When participants estimated the score value and uncertainty with Suppress or ShowSuppress, they gave smaller uncertainty bounds than with Baseline and Show, which is consistent with the lower uncertainty due to suppression. P11 *“(thought) the suppressed subscore will make the wine score more accurate.”* Participants were more decisive to give an estimate further from decision boundary 50, with lower uncertainty. After viewing that the scores with Suppress were higher above 50 than with Baseline, P6, P8, and P11 further increased their estimate.

5.5.4. Preference to Suppress Uncertainty, but Divergent Opinions to Show Suppressed Uncertainty

We found that most participants appreciated the suppressed subscore, and some participants liked to see the subscore uncertainty when subscores are suppressed. P1 preferred Suppress over Show because *“before suppressing, (the Show version) is not very accurate.”* P9 preferred ShowSuppress over Show because *“I don’t really have to pay attention to how much the (input) uncertainty is anymore. I just assume the system already accounts for it to calculate the subscore and final score.”* When comparing between Suppress and ShowSuppress, participants had divergent opinions. P8 thought the hybrid of showing and suppressing is better than suppressing only. He found the subscore in ShowSuppress to be more helpful than those of Suppress because *“now I am given the uncertainty (of subscores), I can engage by how much the discrepancy there is, and I can take that into account in my overall decision of whether to accept or reject the wine.”* However, although he found ShowSuppress more helpful than Suppress, he had a lower trust of ShowSuppress than Suppress *“because now that I’m given the uncertainty of the subscore, I know that the machine has a certain discrepancy, so I can’t fully trust the machine.”* Relatedly, P4 preferred not being shown the subscore uncertainty (i.e., preferred Suppressed over ShowSuppressed) *“because there will be too much information.”*

5.5.5. Opportunities to Improve Interpretability

Although we have found that participants understood and made good use of different versions of explanations, we noticed that they may occasionally misuse the user interface, questioned the system’s uncertainty handling and asked for more information.

Although attribution explanations are popular and intuitive [4, 57, 63], some participants were curious to learn deeper explanations to fully understand the system. With Baseline, P6 found the subscores unhelpful for her decision making because she did not know how each subscore was calculated and what was its relationship between the input value and subscore value. However, she could not articulate what other information would help her. We prompted that a line chart visualization¹³ or rules¹⁴ could provide deeper explanations, but she felt they would be too much information. We avoided mentioning of machine learning concepts, such as multi-factor linear regression, feature vector spaces, and neural networks to avoid jargon with lay users. This is beyond the scope of our study and relates to AI literacy in the general public [56]. Nevertheless, we limited our study to the 2nd level of the Self-Explaining Scorecard [45] to more deeply investigate the impact of uncertainty even at such a low level of explanation.

Similarly, some participants wanted to know how subscore uncertainties (in Show) were calculated. We avoided discussing error propagation with Monte Carlo simulation, hypothetical instances and sampling, since that would require a more advanced understanding of statistics which is not common in lay knowledge. P9 said that it is good to know the subscore uncertainty, but she wanted to know more: *“I don’t know how much the change in reading would lead to the change in the subscore. ... I see that a small variation in alcohol can lead to larger subscore uncertainty, but I don’t know the direct relationship.”* This suggests an interest in counterfactual explanations [59], which is beyond the scope of our study.

As for Suppress and ShowSuppress, we identified two issues: 1) it requires some effort to understand how suppressing works, and 2) participants would like to personalize the extent of suppression. Regarding the learning curve, participants appreciated the suppressing but some of them found it confusing at first. P7, who used Suppress, felt that *“initially it is confusing to me, it*

¹³For example, with partial dependence plots or generalized additive models (GAM).

¹⁴As generated by association rule mining, decision tree learning, etc.

takes me a while to understand the suppression." Within about a minute, P7 switched repeatedly between Baseline and Suppress to understand. This inspired us to implement the hovering interaction to help users perceive the changes in subscores with and without suppression. Regarding the personalized suppression preference, most participants were fine with strong suppression. For an instance where the Alcohol subscore was suppressed to almost zero, P8 found it acceptable, because "*the uncertainty of Alcohol subscore is very high, so I wouldn't use Alcohol in my decision making.*". However, P6, who used ShowSuppress, felt that the subscore should not be suppressed too much (i.e., all the way to zero), and suppressing should be a little or by half because "*I don't know if the subscore suppressing is actually good and accurately evaluate the quality of each component.*"

5.5.6. Clarity and Usability Issues and Fixes

We identified several minor clarity and usability issues that we fixed in the subsequent quantitative study. P4 was confused about why the subscores were so small and believed that they had to be larger to be consistent with the background knowledge. We clarified that the presented cases were borderline rather than typical, so the subscores would tend to be small. P3 also was interested to know how the subscore is calculated and tried to rationalize it by counting the number of positive bars and subtracting the number of negative bars, but not their subscore values. We clarified that the total score is derived from the sum of subscores, not the count of properties. Another participant preferred to see subscores as additive partial attributions from 0 that sum to the total score, rather than an average value and attribution differences from average. Designers can consider this alternative representation in the future. However, we point out that this may limit an intuitive comparison between suppressed and default attributions as deviations from the mean. P1 found that the subscore uncertainty did not add up to the total score uncertainty as a simple sum. We clarified that the subscores should be a sum of squares to get the square of total score uncertainty. Other minor fixes include amplifying the contrast in the highlight colors of the uncertainty numbers, scaling all tornado plots to the same range.

5.6. Quantitative Study: Impact of Showing and Suppressing Attribution Uncertainty on User Performance and Perception

Following the formative qualitative study where we interviewed participants on the usage and rationale of the explanation techniques, we conducted

a summative quantitative study to evaluate how Showing or Suppressing attribution uncertainty affects user decision performance and perception about the AI system and explanation. In this section, we describe our measures, hypotheses, and results. We used the same primary independent variable of explanation techniques in the quantitative study as in the qualitative study.

5.6.1. Random Variables

Besides the Independent Variables and Controlled Variables described in Section 5.2, we tracked several Random Variables in the quantitative study to help calculate dependent variables and to account for potential confounds.

We varied system predicted and actual scores as follows:

1. Trial Sequence (1 to 15): to describe the order of the tasks received by participants. This is only used in the quantitative study.
2. System Score: The system wine quality score, denoted as $Score_{System}$. For Suppress and ShowSuppress which show the suppressed and default subscores, we also track the baseline score, denoted as $Score_{SystemBaseline}$.
3. System Score Uncertainty of Baseline: The uncertainty of system quality score from the Baseline, denoted as $ScoreUncertainty_{SystemBaseline}$.
4. Actual Score: The ground truth quality score, denoted as $Score_{Actual}$.
5. Actual Decision (2 levels): Accept or Reject. Whether the Actual Score is greater than 50 or not. The Actual Decision of half of the selected wine instance in the study is Accept.

Confounds due to user background include:

6. Uncertainty Decisiveness: We ask participants three 7-point Likert scale questions about their ability to make decision uncertainty [30], and this measures whether they were higher or lower than neutral.
7. Uncertainty Tolerance: We ask participants three 7-point Likert scale questions about their uncertainty Tolerance [9].

5.6.2. Dependent Variables

We measured and computed various dependent variables to understand the participants' decision correctness, task time, and confidence in decision making, trust¹⁵ in the system's inference and helpfulness of the system. We

¹⁵We did not evaluate trust appropriateness on the system and controlled the system correctness to be good for all instances because: 1) to learn when it is appropriate to trust, the user will need ground truth feedback to assess system correctness; 2) it takes

Dependent Variable	Definition
Log(Time)	Logarithmic transform of task time per wine decision trial.
Decision Closeness	Negative of difference between the user's score and system's displayed score, i.e., $- Score_{User} - Score_{System} $
Relative Decision Uncertainty	Ratio of user's score uncertainty and system's baseline uncertainty, i.e., $ScoreUncertainty_{User}/ScoreUncertainty_{SystemBaseline}$. A ratio <1 means that the user perceives lower uncertainty than the system.
Decision Quality	Negative of difference between the user's score and actual (ground truth) score, i.e., $- Score_{User} - Score_{Actual} $
Sum(Decision Correctness)	Sum of correct decisions (accept/reject) for all 15 trials per system version.
Trust	7-point Likert scale rating
Confidence	(strongly disagree to strongly agree).
Helpfulness	

Table 1: Dependent variables measured and analyzed in the quantitative user study.

recorded measures of the system prediction and actual (ground truth) scores used to calculate subsequent dependent variables: We further computed dependent variables to more determine the impact on relevant outcomes, such as decision quality. See Table 1 for definitions and calculations.

5.6.3. Hypotheses

We hypothesized how different explanation techniques would influence each dependent variable. See Table 2 middle column. For Decision Time (H1), both Show and Suppress provide additional information¹⁶, so we hypothesized that they will take more time to read than Baseline, and their combination of them ShowSuppress (SS) will take longer than either separately. Since Show has higher uncertainty than Suppress, we hypothesize that

many trials to observe and learn even a simple system failure pattern of a toy system [5]; 3) if ground truth feedback is unknown, users will need strong domain knowledge or technical background to identify system errors, which is not easy even for experts without any social discussion and experimenting, not to mention the lay users. Refer to [38, 53] for some results that users tend to over-trust AI even when they may be inappropriate; this is beyond the scope of our current study. We defer the study of uncertain explanations of inappropriate model behavior to future work.

¹⁶Show adds subscore uncertainty numbers, Suppress displays the UI twice to compare with Baseline

it would make users more uncertain in decision making and would slow them down than Suppress and Baseline. We hypothesize that Decision Closeness (H2), Trust (H6), Confidence (H7) and Helpfulness (H8) will be correlated and be ordered by the most suppressed and transparent (SS) as the best, followed by Suppress, Baseline, and then Show as last, because showing uncertainty can lead to distrust [53]. For Relative Decision Uncertainty (H3), we hypothesized that showing uncertainty (Show, ShowSuppress) would lead to higher reported uncertainty than unshown (Baseline, Suppress, respectively) because of increased awareness about uncertainty; suppressing (Suppress, ShowSuppress) will have lower reported uncertainty due to the uncertainty that is communicated being smaller. We hypothesize that Decision Quality (H4) and Decision Correctness (H5) will be correlated and suppression will lead to better decisions due to the model correctness is controlled as good and the suppressing improved expected faithfulness to model in Section 4.4 (Figure 3), and that showing more information can help the decision to be correct.

Measure	Hypotheses	Results
H1. Decision Time	Baseline < Suppress < Show < SS	Baseline = Suppress = SS < Show
H2. Decision Closeness	Show < Baseline < Suppress ≤ SS	Baseline = Show ≤ SS ≤ Suppress
H3. Relative Decision Uncertainty	Suppress ≤ SS < Baseline < Show	Suppress < SS < Baseline = Show
H4. Decision Quality	Baseline ≤ Show < Suppress ≤ SS	Show = Baseline = SS = Suppress
H5. Decision Correctness	Baseline ≤ Show < Suppress ≤ SS	Show = Baseline = Suppress = SS
H6. Trust	Show < Baseline < Suppress < SS	Show = Baseline < Suppress = SS
H7. Confidence	Show < Baseline < Suppress ≤ SS	Show = Baseline = Suppress = SS
H8. Helpfulness	Show < Baseline < Suppress < SS	Show = Baseline < Suppress = SS

Table 2: Hypotheses and results from quantitative analysis in terms of experiment hypotheses. SS refers to the ShowSuppress technique. Sign “<” indicates a significant difference at $p < .0001$, “≤” indicates marginal difference at $p < .001$, and “=” indicates no significant difference at $p > .01$. Although adjacent comparisons are not significant, other comparisons may be significantly different and this is indicated as branched lines with “<”. Red text and signs indicate results that do not agree with the hypotheses.

We pre-registered our experiment variables, hypotheses, data exclusion criteria and analysis methods at AsPredicted¹⁷ before we collected data. Decision Correctness is an additional dependent variable that we investigated besides what we pre-registered. It is the classification version of Decision Quality, and they have the same hypothesis.

5.6.4. Statistical Analysis and Results

We recruited participants from Amazon Mechanical Turk (MTurk). To ensure the data quality, our survey was only available to Mturk workers with high qualification (at least 5,000 completed HITs with above 97% approval rate). Participants were compensated US\$ 5.00 once they pass the screening quiz and completed the survey in around 35 minutes. Of the 270 MTurk workers who attempted the survey, 147 passed the screening quiz to complete the survey (54.4% pass rate). If a participant made correct decisions for more than 24 instances (out of 30), she would get a bonus of \$0.25 per correct instance. A participant can get up to US\$7.00 payment. To manage MTurk quality issues [43], we had the following pre-registered data exclusion criteria:

1. Per-trial responses with abnormal outlier completion times (too long or too short).
2. Per-trial responses with inconsistent answers (e.g. accept a wine but rating score is below 50)
3. Per-condition responses from participants who failed more than one third pre-condition comprehension questions
4. Responses that are too identical across trials and questions, which suggests participants rushing through and not answering questions carefully.
5. Participants who provide meaningless or nonsensical free text answers

We recruited 147 participants (32.43% female, median age 36 years) who passed the screening quiz and completed the survey but eliminated a few responses not satisfying the exclusion criteria. Ultimately, 131 respondents were included for analysis.

We divided the dependent variables into per-trial and per-condition variables. We fit one multivariate linear mixed effects models for per-trial variables, and another for per-condition variables. Table 3 describes the details of

¹⁷The anonymized pre-registration document is available at <https://aspredicted.org/blind.php?x=ub3mt2>

Response	Linear Mixed Effects Model (+ Participant as random effect)	p>F	R²	f²
Log(Time)	Technique +	<.0001	.530	1.13
	Trial Sequence	<.0001		
Decision Closeness	Technique +	.0003	.495	0.98
	Trial Sequence		<i>n.s.</i>	
Decision Relative Uncertainty	Technique +	<.0001	.649	1.85
	Trial Sequence		<i>n.s.</i>	
Decision Quality	Technique +	<.0001	.594	1.46
	Trial Sequence +		<i>n.s.</i>	
	Actual Score Closeness +	<.0001		
	Technique × Actual Score Closeness		<i>n.s.</i>	
Sum(Decision Correctness)	Technique	<.0001	.714	2.49
Trust	Technique	<.0001	.721	2.59
Confidence	Technique	.0002	.744	2.91
Helpfulness	Technique	<.0001	.524	1.10

Table 3: Statistical analysis of dependent variable responses and factors (one per row), fit as linear mixed effects models, all with Participant as random effect. Per-trial dependent variables were modeled with Technique and Trial Sequence as fixed effects, and per-condition dependent variables were modeled with only Technique as fixed effect. Decision Correctness was modeled as the per-condition metric Sum(Decision Correctness per-trial) to allow more insightful analysis with parametric modeling. Decision Correctness is a binary variable (0 or 1). Decision Quality depends on the actual (ground truth) score of the instance, so we add Actual Score Closeness (negative absolute difference of actual score to decision boundary 50) as fixed effect and Technique × Actual Score Closeness as interaction effect. *n.s.* refers to no significant difference at p>.01. p>F refers to the significance level of an ANOVA for each fixed effect. R² and f² refer to the model’s coefficient of determination and Cohen’s f², respectively. All models have strong effect size with f²>0.35 and R² \gtrsim 0.5 indicating that they explain at least 50% of variance.

the models with fixed, interaction, and random effects. Generally, all models had good to very good fit (high R²).

Due to a large number of comparisons in our analysis, we consider differences with p<.001 as significant and p<.005 as marginally significant. Most significant results are reported as p<.0001. This is stricter than a Bonferroni correction for k = 50 comparisons (significance level= .05/50).

Following advice from Dragicevic [19], we performed a logarithmic transform on time, Log(Time), which is commonly done for time measurements

[41] to correct for positive skewness in time measurements and mitigate timing outliers [67]. We had analyzed task time with a logarithmic transform to be able to fit a linear mixed effects model with Participant as a random effect. To verify the effect without the log-normality assumptions, we performed non-parametric tests on task time to compare the medians between conditions¹⁸. These verified the significant differences found in the parametric model. Note that since the non-parametric tests do not account for the Participant as a random effect, so it does not consider the reduced variance due to individuals with the within-subjects experiment design.

Figure 9 shows the details of statistical analysis results with Technique as the main fixed effect. We describe key insights in terms of practical differences. Note that many of the numbers are in terms of our experiment apparatus (15 trials per condition, with borderline wine instances with high system estimate uncertainty) and the wine rating task (e.g., score from 0 – 100, but likely within 30 – 70). Therefore, we also interpret the results in terms of percent differences to give clarity to how the results could scale to other contexts. We describe findings for each dependent variable:

Task Time: Using Show explanations is slower than Baseline explanations by 3.01s ($M=18.3$ vs. $15.3s$), which is 19.7% slower. Suppress or ShowSuppress explanations do not take significantly more time than Baseline.

Decision Closeness and Decision Quality have units in terms of score difference, for scores that vary from 0 to 100. We relate them to the system decision quality (the negative difference between system baseline score and actual score, i.e., $-|Score_{SystemBaseline} - Score_{Actual}|$, which has Medians -4.01.

Decision Closeness: Using Suppress and ShowSuppress increase Decision Closeness from Baseline -2.92 by 0.350 and 0.328 (11.1% and 9.7%), respectively. In the suppressed conditions, since both suppressed versions and corresponding unsuppressed versions were provided for comparison, we compared the two distances from user score to suppressed and unsuppressed sys-

¹⁸A Kruskal-Wallis H test by ranks found a significant difference between the task Time across Technique ($p<.0001$), and pairwise Wilcoxon Rank Sum tests found that Show technique had higher task Time than all other techniques (all $p<.0001$). The less statistically efficient Median test of the number of points above the median was also significant ($p<.0001$), and pairwise Median tests also found that Show had higher task Time than None, Baseline, Suppress, and ShowSuppress ($p = .0082, <.0001, <.0001, .0067$, respectively).

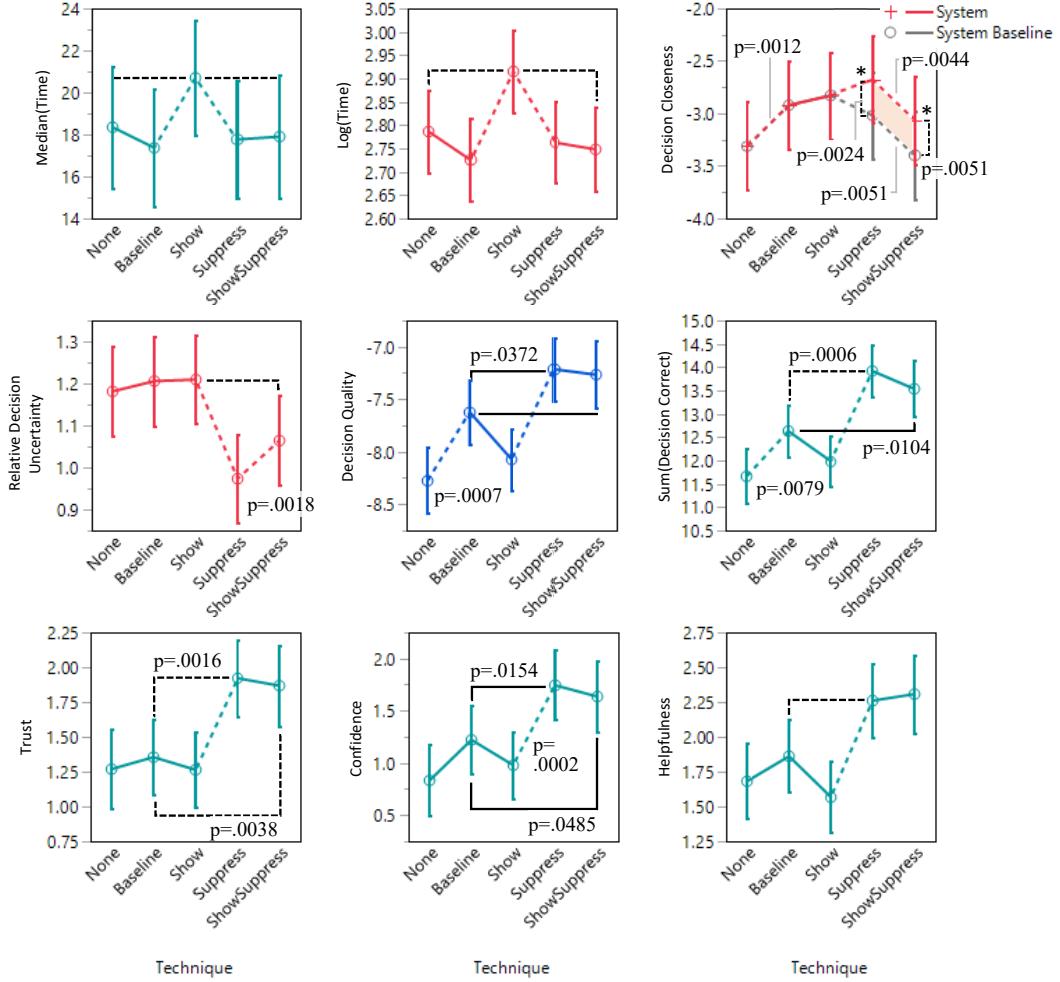


Figure 9: Results showing how explanation types influence 8 dependent variables. Dotted lines indicate extremely significant $p < .0001$ comparisons, otherwise very significant as stated; solid lines indicate no significance at $p > .01$. Error bars indicate a 90% confidence interval. Distance Closeness includes an additional grey line to indicate closeness to the Baseline score that is shown with the suppressed score for suppressed explanation types. Per-trial dependent variables are shown as red lines, and per-system dependent variables are shown as teal lines. Decision Quality was modeled with Actual Score Ambiguity (how close to decision boundary 50) as a fixed effect, and this is shown as a blue line.

tem scores by ratio: $|Score_{User} - Score_{System}| / |Score_{User} - Score_{SystemBaseline}|$, and the median is 0.889 (for 67.5% of participants the distance ratio is smaller

than 1), indicating that participants tend to choose a score closer to the suppressed score than the baseline score when in any suppressed condition, and there was no significant difference between ShowSuppress and Suppress. Using Show explanations do not change Decision Closeness. For Suppress and ShowSuppress, Distance Closeness to System is marginally higher than Distance Closeness to Baseline, suggesting that participants did prefer to follow the suppressed score (System) than the unsuppressed one (Baseline).

Decision Quality: Using any explanation improves Decision Quality compared to None. Using Suppress or ShowSuppress increases Decision Quality more than using Show from -8.08 by 0.860 and 0.810 (10.7% and 10.0%), respectively. However, Show, Suppress and ShowSuppress are not significantly different from Baseline.

Relative Decision Uncertainty: Using Suppress and ShowSuppress decreases user perceived Relative Decision Uncertainty from Baseline at 1.21 by 0.232 and 0.142 (19.3% and 11.7%), respectively. When using explanations without suppression (None, Baseline, Show), users perceived higher decision uncertainty than what the system showed, i.e., >1. Relative Decision Uncertainty for Suppress is significantly lower than ShowSuppress ($M=0.97$ vs. 1.06).

Sum(Decision Correct): Using Suppress and ShowSuppress increases decision correctness compared to Show from 12.0 correct answers by 1.94 and 1.56 (16.2% and 13.0%), respectively. Only Suppress significantly improves decision correctness compared to Baseline, and from 12.6 correct answers by 1.29 (10.2%). Show explanations do not improve decision correctness compared to Baseline.

Trust: Using Suppress and ShowSuppress increases Trust rating compared to Baseline from 1.35 (above 1=“Slightly Agree”) to 1.92 and 1.87 (both almost 2=“Agree”). Show does not improve Trust compared to Baseline.

Confidence: Using Suppress increases Confidence rating compared to Baseline from 1.22 (above 1=“Slightly Agree”) to 1.74 (towards 2=“Agree”). Show and ShowSuppress do not improve Trust compared to Baseline.

Helpfulness: Using Suppress and ShowSuppress increases Helpfulness rating compared to Baseline from 1.86 (below 2=“Agree”) to 2.26 and 2.30 (above 2=“Agree”). Show does not improve Trust compared to Baseline.

5.6.5. Interpreting Results With Respect to Hypotheses

Next, we interpret our results in terms of our hypotheses of how different explanation techniques could be beneficial or detrimental to user performance and opinion. Table 2 compares the results with the hypotheses and highlights the differences. The key insights are:

1. ShowSuppress is not the slowest to use, but Show is instead, suggesting that the trust and confidence gained due to suppressing uncertainty helped participants to be more decisive. Indeed, participants are no slower using ShowSuppress than Baseline or Suppress.
2. Though we expected the Show technique to have the lowest Decision Closeness, Trust, Confidence, and Helpfulness because it reveals model uncertainty, it is not significantly worse than Baseline.
3. Suppress and ShowSuppress do reduce perceived Decision Uncertainty, improve Trust and Helpfulness compared to Baseline, its impact on Decision Quality, Decision Correctness and Confidence is less significant.

In summary, showing attribution uncertainty that is large may not improve user performance and perception, and yet costs more task time. Therefore, it is more important to mitigate uncertainty first. Showing the uncertainty suppression achieves the same performance and perception improvement as suppressing uncertainty compared to baseline, and does not suffer from the detrimental effects of showing unsuppressed attribution uncertainty.

5.6.6. Factor Analysis of Showing and Suppressing

The four explanation Techniques (excluding None) can also be considered in terms of the two factors Showing and Suppressing. For technique Show and ShowSuppress, factor Showing is true; for technique Baseline and Suppress, factor Showing is false. For technique Suppress and ShowSuppress, factor Suppressing is true; for technique Baseline and Show, factor Suppressing is false. We performed contrast t-tests with fixed effects Showing, Suppressing, and contrast interaction effect Showing \times Suppressing. Table 4 shows the results of contrast t-tests and 2-way ANOVA on these factors. There are significant fixed and interaction effects for Log(Time) as previously seen in Figure 9. All explanation Techniques achieve indistinguishable levels of Decision Closeness. For all other dependent variables, there is no effect of Showing attribution uncertainty and no interaction effects, but there are significant effects Suppressing attribution uncertainty.

Response	Fixed Effects ($p > F$)		
	Showing	Suppressing	Showing \times Suppressing
Log(Time)	<.0001	.0014	<.0001
Decision Closeness	n.s.	n.s.	n.s.
Decision Closeness (to Baseline)	n.s.	.0004	n.s.
Relative Decision Uncertainty	n.s.	<.0001	n.s.
Decision Quality	n.s.	<.0001	n.s.
Sum(Decision Correctness)	n.s.	<.0001	n.s.
Trust	n.s.	<.0001	n.s.
Confidence	n.s.	<.0001	n.s.
Helpfulness	n.s.	<.0001	n.s.

Table 4: Results of t-tests and 2-way ANOVA on the factors of the explanation techniques for Showing or Suppressing attribution uncertainty in explanations.

5.6.7. Some Differences Due to Uncertainty Intolerance

Response	Linear Mixed Effects Model (+ Participant as random effect)	p>F	R ²	f ²
Decision Closeness	Technique +	.0002	.496	0.99
	Trial Sequence +	n.s.		
	Uncertainty Tolerant +	n.s.		
	Uncertainty Decisive +	n.s.		
	Technique \times Uncertainty Tolerant +	<.0001		
	Technique \times Uncertainty Decisive	n.s.		
Confidence	Technique +	<.0001	.744	2.91
	Uncertainty Tolerant +	n.s.		
	Uncertainty Decisive +	n.s.		
	Technique \times Uncertainty Tolerant +	.0032		
	Technique \times Uncertainty Decisive	n.s.		

Table 5: Statistical analyses including user uncertainty intolerance and decisiveness as linear mixed effects models with fixed and interaction effects (one per row), and with Participant as a random effect. n.s. refers to no significant difference at $p>.01$. p>F refers to the significance level of an ANOVA for each fixed effect. R² and f² refer to the model's coefficient of determination and Cohen's f², respectively. All models have a strong effect size with $f^2>0.35$ and $R^2 \gtrsim 0.5$ indicating that they explain at least 50% of the variance.

As an additional analysis, we investigated whether differences in indi-

vidual risk tolerance influenced their perception and use of the explanation techniques. We note that the survey questions on personality types are context-free and prone to be inaccurate to characterize real-world intolerance perception. Nevertheless, this exploratory analysis can provide some formative insights. We found that the six survey questions on risk tolerance were correlated and performed the common factor analysis using maximum likelihood with varimax rotation to group responses into two factors — uncertainty tolerance (first 3 questions) and uncertainty decisiveness (last 3 questions), see Figure C.7. The final number of factors are statistically significant by the Bartlett Test of Sphericity (all $p < .0001$) and explains 66.5% of the response variance. To support the analysis of interaction effects, we further binarized the factors into whether the response is above the median to derive the fixed effects Uncertainty Tolerant and Uncertainty Decisive, respectively. We fit a multivariate linear mixed effects model as described in Table 5.

We found several significant interaction effects for per-trial Decision Closeness and per-condition Confidence (see Figure 10). We performed contrast t-tests for specific differences identified. When viewing explanations that Show attribution uncertainty, participants who were more Uncertainty Tolerant had higher Decision Closeness (agreement with the AI system) than those who were less tolerant ($M = -2.11$ vs. -3.58 , $p = .0007$). This suggests that uncertainty intolerant users are less trusting of AI when they see that the AI is uncertain. When viewing Baseline explanations that do not show or suppress attribution uncertainty, Confidence in decision making is higher for more Uncertainty Tolerant users than less tolerant ones ($M = 1.81$ vs. 0.64 , $p = .0006$). This suggests that uncertainty tolerant users become more confident when viewing the AI's explanation, though without uncertainty in its explanations; Baseline explanations do not raise the confidence of uncertainty intolerant users. However, for uncertainty intolerant users, they had significantly higher confidence using Suppress compared to Baseline ($M = 1.86$ vs. 0.64 , $p = .0001$), suggesting that the suppression helped make them more decisive.

5.7. Summary of Qualitative and Quantitative Results

We summarize key findings unified from our qualitative and quantitative user studies. Compared to baseline explanations, showing attribution uncertainty helps users to be more aware of and understand internal attribution uncertainty and, in general, does not increase model (score output)

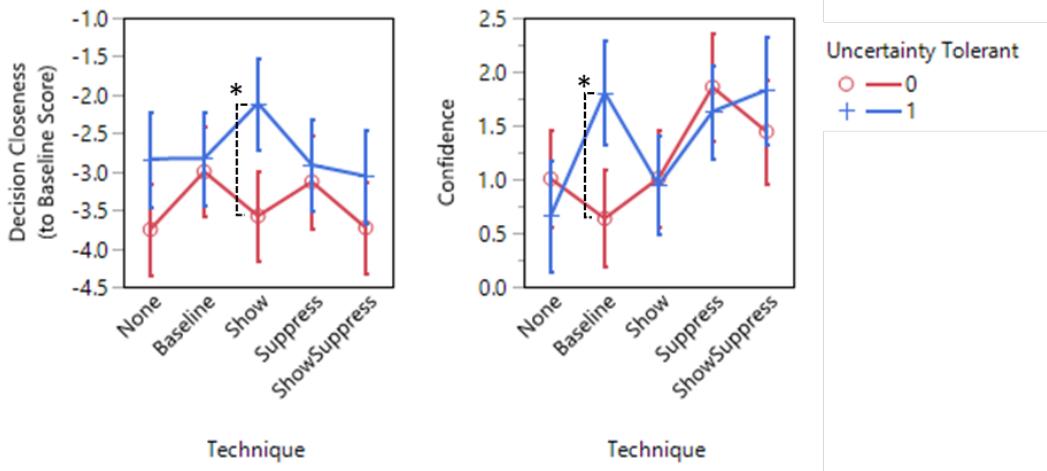


Figure 10: Results showing how participants who are more Uncertainty Tolerant or less tolerant are differently influenced by different explanation types. Significant contrast t-test effects are indicated with * and dotted line. Error bars indicate a 90% confidence interval.

uncertainty or decrease decision confidence, though it requires more decision time. However, this improvement in understanding did not improve user decision correctness or tendency to copy system predictions (decision closeness). Although, on average, user trust in and perceived helpfulness system predictions with shown attribution uncertainty was the same as for baseline explanations, our interviews revealed divergent opinions regarding trust and helpfulness. From interviews, we learned that users who rated more positively appreciated understanding the relationship between input uncertainty and output uncertainty, while users who rated less positively were concerned about knowing more sources of high uncertainty. We also saw a similar divergence in our quantitative analysis that shown attribution uncertainty, users who were more uncertainty tolerant were more likely to copy the system score than users who were less tolerant. This suggests that uncertainty intolerance can lead to more distrust in predictions when more details of uncertainty are shown.

Compared to baseline explanations, suppressing attribution uncertainty helps to improve user decision quality, correctness, and confidence, without costing more decision time. Users found it more helpful and trustworthy too, because of the visibly reduced model (output) uncertainty and reduced

concern regarding uncertain inputs due to their smaller attribution. Many users were appreciative of the "compensation" to re-allocate attributions due to the suppression. With suppressed attributions, showing the attribution uncertainty (ShowSuppress) did not suffer from performance and perception issues of show unsuppressed uncertainty, because users were less worried about the small uncertainties. Thus, showing suppressed attributions with uncertainty did not even cost more decision time than using baseline explanations. Finally, through interviews, we found interest in personalizing the extent of suppression to see a more subdued effect or to customize based on applications.

6. Design Implications: Show or Suppress?

In this work, we have investigated uncertainty that is propagated into feature attributions and compared two alternative approaches of showing uncertainty within explanations or suppressing attributions to make explanations less reliant on uncertain input. We note that there are many needs for a variety of explanations and deeper mechanistic explanations as demonstrated in our interview results and in literature [12, 17, 26, 50, 63]. Our approaches are complementary to these explanations, but further study is needed to investigate the impact of showing or suppressing uncertainty in those explanations. Here, we discuss recommendations for whether to show or suppress attribution explanation uncertainty in explanations based on findings from our simulation experiments and user studies.

6.1. Communicating Explanation Attribution Uncertainty

Much research to visualize uncertainty has focused on helping users to identify uncertain inputs [13, 58], but this requires users to apply their domain expertise to understand how they relate to the analysis task. With machine learning models or other automation systems, research on communicating uncertainty has focused on the model, inference or outcome uncertainty, typically expressed as model confidence [39, 53, 77]. However, these show an overall uncertainty due to a single decision and does not provide additional details for the source of uncertainties. Communicating both input uncertainty and model uncertainty can provide more insights to users, but the relationship between input and model uncertainty remains unclear. In this work, we have proposed and demonstrated the benefits and issues of communicating uncertainty *within* an explanation model to provide more

transparency uncertainty. We found that users could understand the model better by being able to trace how uncertainty propagates from input, to attributions, to model output. Therefore, attribution uncertainty can provide additional insight into the internal workings of AI models.

6.2. Show Attribution Uncertainty When Suppressed

A key recommendation from our results is to show attribution uncertainty when it is suppressed to be low. Otherwise showing high attribution uncertainty would worry the users, increase the uncertainty of their decision, reduce the confidence in their decisions, slow down their decision making, and ultimately limit their trust towards the system. Our findings are similar to Lim and Dey [53] who found that when model uncertainty is low, showing explanations improves user trust, but when model uncertainty is high, showing explanations hurts user trust. In our work, instead of requiring low uncertainty, our proposed Suppress technique can reduce model uncertainty to low enough levels and achieve the gains in perceived helpfulness and trust towards the system. Though, we recommend being judicious in how much suppression to exercise, and to consider allowing users to interactively control the suppression level.

We note that with ShowSuppress explanations, some users may be misled to think that some input features would not affect the predictor model output much. We highlight that baseline (Show) and regularized (Show-Suppress) explanations serve different purposes. Baseline explanations aim to faithfully represent the Predictor for each instance even with the noise ϵ , i.e., $g_f(x + \epsilon) \approx f(x + \epsilon)$, and provide representative feature weights of how the Predictor thinks. In contrast, Regularized Explainers prioritize a more robust explanation to discount the effect of uncertain features, i.e., $\tilde{g}_f(x + \epsilon) \approx f(x)$, and provide an interpretation that defers influences to more reliable features. Under uncertainty, even though the Predictor could produce an ambiguous prediction, ShowSuppress would inform the user with a more confident explanation that we have shown to be more accurate on average. To ameliorate the confusion between the two explanation types, users should be reminded of the distinction between Show and ShowSuppress explanations, i.e., $g_f(x + \epsilon) \neq \tilde{g}_f(x + \epsilon)$. They should be shown together as we have done, with Show as the primary explanation and ShowSuppress as supplementary. Providing explanations from a Regularized Predictor will avoid this issue, but would require Predictor models to be retrained. Furthermore, while ShowSuppress explanations increased decision quality for

correct system predictions, they may not increase decision quality for wrong predictions, since users may not be able to perceive uncertainty to support counterfactual reasoning. Similar to prior works [5, 38, 53], we expect that ShowSuppress explanations would also lead to inappropriate trust for wrong predictions and defer a fuller investigation to future work.

6.3. Alternative Visualizations of Attribution Uncertainty

Among the many ways to visualize uncertainty [34, 39], in this work, we have employed the violin plot and confidence interval (\pm number) to show attribution uncertainty. On the one hand, we demonstrated augmenting a tornado plot with many violin plots to visualize details of each uncertainty distribution. This is informative for savvy users with good technical, statistical, and graph literacy. On the other hand, we evaluated the confidence interval number with lay users online to ensure simple, quick interpretation, since this does not directly require graph literacy or familiarity with probability distributions. In pilot studies, with participants screened for graph literacy, we found that users of violin-tornado plots had higher trust in the system than users of baseline tornado plots. This discrepancy from our current results could be because these users could appreciate the mathematical detail provided in these explanations and found them more trustworthy. Future work can investigate showing attribution uncertainty to different target users with different user-friendly uncertainty visualization methods, such as quantile dot plots [39] or hypothetical outcome plots (HOP) [34].

7. Conclusion

We have highlighted and investigated how attribution explanation uncertainty can impact explanations in machine learning, particularly in terms of user trust, confidence, and decision-making. Informed by different uncertainty coping strategies, we proposed two techniques to manage attribution explanation uncertainty: showing uncertainty in attribution explanations and suppressing attribution uncertainty by reducing and reallocating attributions. We conducted a simulation study, qualitative interviews, and quantitative user evaluation to compare these techniques with baseline explanations. We found that showing attribution uncertainty helps with user understanding of the prediction models, but does not improve trust, confidence and decision making, yet slows down decision making; suppressing uncertainty reduces decision uncertainty, improves trust towards prediction models, user

confidence in decision making, and decision quality. This demonstrates the importance to carefully communicate uncertainty *within* attribution explanations to improve the trustworthiness of artificial intelligence systems.

8. Acknowledgement

We thank Ashraf Abdul, Dr. Lyu Yan, Dr. Zhang Yehong and Dr. Wang Shengyu for their assistance in the discussion, piloting study, providing language help and proof-reading. We thank all our participants for their dedication of time and precious feedback. This work was supported in part by the Ministry of Education, Singapore and was carried out at the NUS Institute for Health Innovation and Technology (iHealthtech) under grants T1 251RES1804 and R-722-000-004-731.

References

- [1] Abdelaziz, A.H., Watanabe, S., Hershey, J.R., Vincent, E., Kolossa, D., 2015. Uncertainty propagation through deep neural networks, in: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, pp. 3561–3565.
- [2] Abdul, A., Vermeulen, J., Wang, D., Lim, B.Y., Kankanhalli, M., 2018. Trends and trajectories for explainable, accountable and intelligible systems: An HCI research agenda, in: Conference on Human Factors in Computing Systems - Proceedings. doi:10.1145/3173574.3174156.
- [3] Antifakos, S., Kern, N., Schiele, B., Schwaninger, A., 2005. Towards improving trust in context-aware systems by displaying system confidence, in: ACM International Conference Proceeding Series, ACM Press, New York, New York, USA. pp. 9–14. doi:10.1145/1085777.1085780.
- [4] Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., Samek, W., 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one 10, e0130140.
- [5] Bansal, G., Nushi, B., Kamar, E., Lasecki, W.S., Weld, D.S., Horvitz, E., 2019. Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance. Proceedings of the AAAI Conference on Human Computation and Crowdsourcing 7, 2–11. URL: <https://www.aaai.org/ojs/index.php/HCOMP/article/view/5285>.
- [6] Bica, M., Demuth, J.L., Dykes, J.E., Palen, L., 2019. Communicating hurricane risks: Multi-method examination of risk imagery diffusion, in: Conference on Human Factors in Computing Systems - Proceedings, Glasgow. doi:10.1145/3290605.3300545.
- [7] Boukhelifa, N., Perrin, M.E., Huron, S., Eagan, J., 2017. How data workers cope with uncertainty: A task characterisation study, in: Conference on Human Factors in Computing Systems - Proceedings, pp. 3645–3656. doi:10.1145/3025453.3025738.
- [8] Cai, C.J., Reif, E., Hegde, N., Hipp, J., Kim, B., Smilkov, D., Wattenberg, M., Viegas, F., Corrado, G.S., Stumpe, M.C., Terry, M., 2019.

Human-centered tools for coping with imperfect algorithms during medical decision-making, in: Conference on Human Factors in Computing Systems - Proceedings. doi:10.1145/3290605.3300234.

- [9] Carleton, R.N., Norton, M.A.J., Asmundson, G.J., 2007. Fearing the unknown: A short version of the Intolerance of Uncertainty Scale. *Journal of Anxiety Disorders* 21, 105–117. doi:10.1016/j.janxdis.2006.03.014.
- [10] Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., Elhadad, N., 2015. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission, in: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1721–1730. doi:10.1145/2783258.2788613.
- [11] Chen, J., Wu, X., Rastogi, V., Liang, Y., Jha, S., 2019. Robust Attribution Regularization, in: Advances in Neural Information Processing Systems, pp. 14300–14310.
- [12] Cheng, H.F., Wang, R., Zhang, Z., O’Connell, F., Gray, T., Harper, F.M., Zhu, H., 2019. Explaining decision-making algorithms through UI: Strategies to help non-expert stakeholders, in: Conference on Human Factors in Computing Systems - Proceedings. doi:10.1145/3290605.3300789.
- [13] Correa, C.D., Chan, Y.H., Kwan-Liu, M., 2009. A framework for uncertainty-aware visual analytics, in: VAST 09 - IEEE Symposium on Visual Analytics Science and Technology, Proceedings, pp. 51–58. doi:10.1109/VAST.2009.5332611.
- [14] Cortez, P., Cerdeira, A., Almeida, F., Matos, T., Reis, J., 2009. Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems* 47, 547–553. doi:10.1016/j.dss.2009.05.016.
- [15] Czarnecki, W.M., Podolak, I.T., 2013. Machine learning with known input data uncertainty measure, in: IFIP International Conference on Computer Information Systems and Industrial Management, Springer. pp. 379–388.

- [16] Datta, A., Sen, S., Zick, Y., 2016. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems, in: 2016 IEEE symposium on security and privacy (SP), IEEE. pp. 598–617.
- [17] Dodge, J., Vera Liao, Q., Zhang, Y., Bellamy, R.K., Dugan, C., 2019. Explaining models: An empirical study of how explanations impact fairness judgment, in: International Conference on Intelligent User Interfaces, Proceedings IUI, ACM. pp. 275–285. URL: <https://doi.org/10.1145/3301275.3302310>, doi:10.1145/3301275.3302310.
- [18] Doshi-Velez, F., Kim, B., 2017. Towards a rigorous science of interpretable machine learning. *stat* 1050, 2.
- [19] Dragicevic, P., 2016. Fair statistical communication in hci, in: Modern statistical methods for HCI. Springer, pp. 291–330.
- [20] Eiband, M., Schneider, H., Bilandzic, M., Fazekas-Con, J., Haug, M., Hussmann, H., 2018. Bringing transparency design into practice, in: International Conference on Intelligent User Interfaces, Proceedings IUI, ACM. pp. 211–223. doi:10.1145/3172944.3172961.
- [21] Ellsberg, D., 1961. Risk, Ambiguity, and the Savage Axioms. *The Quarterly Journal of Economics* 75, 643. doi:10.2307/1884324.
- [22] Erion, G., Janizek, J.D., Sturmefels, P., Lundberg, S., Lee, S.I., 2019. Learning explainable models using attribution priors. arXiv preprint arXiv:1906.10670 URL: <https://arxiv.org/abs/1906.10670>.
- [23] Eschenbach, T.G., 1992. Spiderplots versus Tornado Diagrams for Sensitivity Analysis. *Interfaces* 22, 40–46. doi:10.1287/inte.22.6.40.
- [24] Fernandes, M., Walls, L., Munson, S., Hullman, J., Kay, M., 2018. Uncertainty displays using quantile dotplots or CDFs improve transit decision-making, in: Conference on Human Factors in Computing Systems - Proceedings. doi:10.1145/3173574.3173718.
- [25] Friedman, N., Geiger, D., Goldszmit, M., 1997. Bayesian Network Classifiers Overfitting and Underfitting With Machine Learning Algorithms. *Machine Learning* 29, 131–163.

URL: <http://link.springer.com/10.1023/A:1007465528199>,
doi:10.1023/a:1007465528199, arXiv:0507464v2.

- [26] Ghai, B., Liao, Q.V., Zhang, Y., Bellamy, R., Mueller, K., 2020. Explainable active learning (xal): An empirical study of how local explanations impact annotator experience. arXiv preprint arXiv:2001.09219 .
- [27] Ghorbani, A., Abid, A., Zou, J., 2019. Interpretation of Neural Networks Is Fragile. Proceedings of the AAAI Conference on Artificial Intelligence 33, 3681–3688. doi:10.1609/aaai.v33i01.33013681.
- [28] Goldstein, D.G., Rothschild, D., 2014. Lay understanding of probability distributions. Judgment and Decision Making 9, 1–14.
- [29] Gosiewska, A., Biecek, P., 2019. Do not trust additive explanations. arXiv preprint arXiv:1903.11420 URL: <https://arxiv.org/abs/1903.11420>.
- [30] Greco, V., Roger, D., 2001. Coping with uncertainty: The construction and validation of a new measure. Personality and Individual Differences 31, 519–534. doi:10.1016/S0191-8869(00)00156-2.
- [31] Guidotti, R., Ruggieri, S., 2019. On the Stability of Interpretable Models, in: Proceedings of the International Joint Conference on Neural Networks, pp. 2976–2987. doi:10.1109/IJCNN.2019.8852158.
- [32] Hintze, J.L., Nelson, R.D., 1998. Violin plots: A box plot-density trace synergism. American Statistician 52, 181–184. doi:10.1080/00031305.1998.10480559.
- [33] Hohman, F., Head, A., Caruana, R., DeLine, R., Drucker, S.M., 2019. Gamut: A design probe to understand how data scientists understand machine learning models, in: Conference on Human Factors in Computing Systems - Proceedings, ACM. p. 13. doi:10.1145/3290605.3300809.
- [34] Hullman, J., Resnick, P., Adar, E., 2015. Hypothetical outcome plots outperform error bars and violin plots for inferences about reliability of variable ordering. PloS one 10, 1–23.

- [35] Jung, M.F., Sirkin, D., Gür, T.M., Steinert, M., 2015. Displayed uncertainty improves driving experience and behavior: The case of range anxiety in an electric car, in: Conference on Human Factors in Computing Systems - Proceedings, pp. 2201–2210. doi:10.1145/2702123.2702479.
- [36] Kahneman, D., Tversky, A., 2013. Prospect theory: An analysis of decision under risk, in: Handbook of the fundamentals of financial decision making: Part I. World Scientific, pp. 99–127.
- [37] Kale, A., Nguyen, F., Kay, M., Hullman, J., 2019. Hypothetical Outcome Plots Help Untrained Observers Judge Trends in Ambiguous Data. *IEEE Transactions on Visualization and Computer Graphics* 25, 892–902. doi:10.1109/TVCG.2018.2864909.
- [38] Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., Wortman Vaughan, J., 2020. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning, in: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, ACM, New York, NY, USA. pp. 1–14. URL: <https://dl.acm.org/doi/10.1145/3313831.3376219>, doi:10.1145/3313831.3376219.
- [39] Kay, M., Kola, T., Hullman, J.R., Munson, S.A., 2016. When (ish) is my bus? User-centered visualizations of uncertainty in everyday, mobile predictive systems, in: Conference on Human Factors in Computing Systems - Proceedings, pp. 5092–5103. doi:10.1145/2858036.2858558.
- [40] Kay, M., Morris, D., Schraefel, M., Kientz, J.A., 2013. There's no such thing as gaining a pound: Reconsidering the bathroom scale user interface, in: UbiComp 2013 - Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing, pp. 401–410. doi:10.1145/2493432.2493456.
- [41] Keene, O.N., 1995. The log transformation is special. *Statistics in Medicine* 14, 811–819. URL: <https://pubmed.ncbi.nlm.nih.gov/7644861/>, doi:10.1002/sim.4780140810.
- [42] Kendall, A., Gal, Y., 2017. What uncertainties do we need in Bayesian deep learning for computer vision?, in: Advances in Neural Information Processing Systems, pp. 5575–5585.

- [43] Kennedy, R., Clifford, S., Burleigh, T., Jewell, R., Waggoner, P., 2018. The Shape of and Solutions to the MTurk Quality Crisis. SSRN Electronic Journal doi:10.2139/ssrn.3272468.
- [44] Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., et al., 2018. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav), in: International conference on machine learning, PMLR. pp. 2668–2677.
- [45] Klein, G., Hoffman, R.R., Mueller, S.T., 2020. Scorecard for Self-Explaining Capabilities of AI Systems. Technical Report. Explainable AI Program, DARPA, Washington, DC.
- [46] Klein, J., Colot, O., 2010. Automatic discounting rate computation using a dissent criterion.
- [47] Koh, P.W., Liang, P., 2017. Understanding black-box predictions via influence functions, in: Proceedings of the 34th International Conference on Machine Learning-Volume 70, JMLR. org. pp. 1885–1894.
- [48] Krause, J., Perer, A., Ng, K., 2016. Interacting with predictions: Visual inspection of black-box machine learning models, in: Conference on Human Factors in Computing Systems - Proceedings, pp. 5686–5697. doi:10.1145/2858036.2858529.
- [49] Kulesza, T., Burnett, M., Wong, W.K., Stumpf, S., 2015. Principles of Explanatory Debugging to personalize interactive machine learning, in: International Conference on Intelligent User Interfaces, Proceedings IUI, pp. 126–137. doi:10.1145/2678025.2701399.
- [50] Lai, V., Liu, H., Tan, C., 2020. ” why is’ chicago’deceptive?” towards building model-driven tutorials for humans, in: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pp. 1–13.
- [51] Letham, B., Rudin, C., McCormick, T.H., Madigan, D., 2015. Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. Annals of Applied Statistics 9, 1350–1371. doi:10.1214/15-AOAS848.
- [52] Lim, B.Y., Dey, A.K., 2011a. Design of an intelligible mobile context-aware application, in: Mobile HCI 2011 - 13th International Conference

on Human-Computer Interaction with Mobile Devices and Services, pp. 157–166. doi:10.1145/2037373.2037399.

- [53] Lim, B.Y., Dey, A.K., 2011b. Investigating intelligibility for uncertain context-aware applications, in: UbiComp’11 - Proceedings of the 2011 ACM Conference on Ubiquitous Computing, pp. 415–424. doi:10.1145/2030112.2030168.
- [54] Lipshitz, R., Strauss, O., 1997. Coping with uncertainty: A naturalistic decision-making analysis. *Organizational Behavior and Human Decision Processes* 69, 149–163. doi:10.1006/obhd.1997.2679.
- [55] Lipton, Z.C., 2018. The mythos of model interpretability. *Communications of the ACM* 61, 35–43. doi:10.1145/3233231.
- [56] Long, D., Magerko, B., 2020. What is AI Literacy? Competencies and Design Considerations, in: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pp. 1–16.
- [57] Lundberg, S.M., Lee, S.I., 2017. A unified approach to interpreting model predictions, in: Advances in Neural Information Processing Systems, pp. 4766–4775.
- [58] McCurdy, N., Gerdes, J., Meyer, M., 2019. A framework for externalizing implicit error using visualization. *IEEE Transactions on Visualization and Computer Graphics* 25, 925–935. doi:10.1109/TVCG.2018.2864913.
- [59] Miller, T., 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267, 1–38.
- [60] Mummolo, J., Peterson, E., 2019. Demand effects in survey experiments: An empirical assessment. *American Political Science Review* 113, 517–529.
- [61] Nielsen, J., Clemmensen, T., Yssing, C., 2002. Getting access to what goes on in people’s heads? reflections on the think-aloud technique, in: Proceedings of the second Nordic conference on Human-computer interaction, pp. 101–110.

- [62] Pallant, J., 2020. SPSS survival manual: A step by step guide to data analysis using IBM SPSS. Routledge.
- [63] Ribeiro, M.T., Singh, S., Guestrin, C., 2016. "Why should i trust you?" Explaining the predictions of any classifier, in: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144. doi:10.1145/2939672.2939778.
- [64] Ross, A.S., Hughes, M.C., Doshi-Velez, F., 2017. Right for the right reasons: Training differentiable models by constraining their explanations. IJCAI International Joint Conference on Artificial Intelligence , 2662–2670doi:10.24963/ijcai.2017/371.
- [65] Rukzio, E., Hamard, J., Noda, C., De Luca, A., 2006. Visualization of uncertainty in context aware mobile applications, in: ACM International Conference Proceeding Series, ACM Press, New York, New York, USA. pp. 247–250. doi:10.1145/1152215.1152267.
- [66] Sacha, D., Senaratne, H., Kwon, B.C., Ellis, G., Keim, D.A., 2016. The Role of Uncertainty, Awareness, and Trust in Visual Analytics. IEEE Transactions on Visualization and Computer Graphics 22, 240–249. doi:10.1109/TVCG.2015.2467591.
- [67] Sauro, J., Lewis, J.R., 2010. Average task times in usability tests: What to report?, in: Conference on Human Factors in Computing Systems - Proceedings, ACM Press, New York, New York, USA. pp. 2347–2350. URL: <http://portal.acm.org/citation.cfm?doid=1753326.1753679>. doi:10.1145/1753326.1753679.
- [68] Schubert, J., 2011. Conflict management in Dempster–Shafer theory using the degree of falsity. International Journal of Approximate Reasoning 52, 449–460.
- [69] Shafer, G., 1976. A mathematical theory of evidence. volume 42. Princeton university press.
- [70] Simpkin, A.L., Schwartzstein, R.M., 2016. Tolerating uncertainty—the next medical revolution? New England Journal of Medicine 375.
- [71] Singh, M., Kumari, N., Mangla, P., Sinha, A., Balasubramanian, V.N., Krishnamurthy, B., 2019. On the benefits of attributional robustness.

arXiv preprint arXiv:1911.13073 URL: <https://arxiv.org/abs/1911.13073>.

- [72] Sundararajan, M., Taly, A., Yan, Q., 2017. Axiomatic attribution for deep networks, in: 34th International Conference on Machine Learning, ICML 2017, pp. 5109–5118.
- [73] Wang, D., Yang, Q., Abdul, A., Lim, B.Y., 2019. Designing theory-driven user-centric explainable AI, in: Conference on Human Factors in Computing Systems - Proceedings. doi:[10.1145/3290605.3300831](https://doi.org/10.1145/3290605.3300831).
- [74] Wppard, K.K., 2012. Thinking, Fast and Slow. *Development and Learning in Organizations: An International Journal* 26, 38–39. doi:[10.1108/14777281211249969](https://doi.org/10.1108/14777281211249969).
- [75] Wu, M., Hughes, M.C., Parbhoo, S., Zazzi, M., Roth, V., Doshi-Velez, F., 2018. Beyond sparsity: Tree regularization of deep models for interpretability, in: 32nd AAAI Conference on Artificial Intelligence, AAAI 2018, pp. 1670–1678.
- [76] Yang, Y., Han, D., Han, C., 2013. Discounted combination of unreliable evidence using degree of disagreement. *International Journal of Approximate Reasoning* 54, 1197–1216. doi:[10.1016/j.ijar.2013.04.002](https://doi.org/10.1016/j.ijar.2013.04.002).
- [77] Yin, M., Vaughan, J.W., Wallach, H., 2019. Understanding the effect of accuracy on trust in machine learning models, in: Conference on Human Factors in Computing Systems - Proceedings. doi:[10.1145/3290605.3300509](https://doi.org/10.1145/3290605.3300509).
- [78] Zafar, M.R., Khan, N.M., 2019. Dlime: A deterministic local interpretable model-agnostic explanations approach for computer-aided diagnosis systems. arXiv preprint arXiv:1906.10263 .
- [79] Zhang, Y., Song, K., Sun, Y., Tan, S., Udell, M., 2019. Why Should You Trust My Explanation? Understanding Uncertainty in LIME Explanations. arXiv preprint arXiv:1904.12991 URL: <http://arxiv.org/abs/1904.12991>.

Appendix A. Proof of Baseline Explainer's Expected Faithfulness

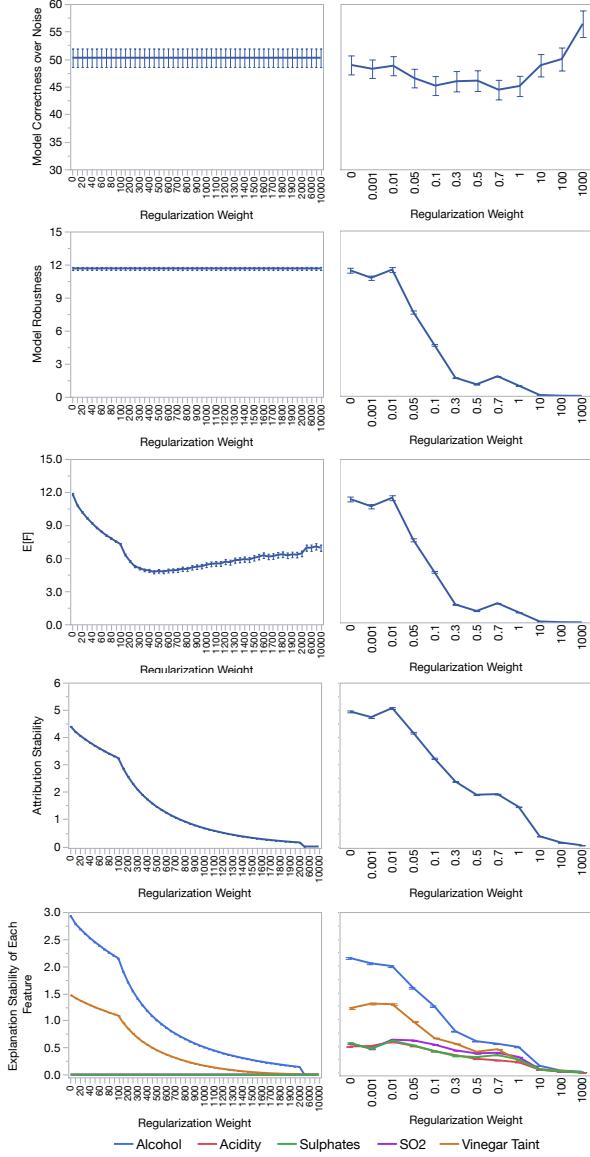
Here we prove that the expected faithfulness distance $E[F_{g_f}]$ of baseline LIME explainer g_f over noisy input $x + \epsilon$ is the same as or worse than point-estimated baseline faithfulness distance F_0 .

Proof. $E_\epsilon[F_{g_f}] \geq F_0$.

$$\begin{aligned}
F_{g_f} - F_0 &= (g_f(x + \epsilon) - f(x))^2 - (g_f(x) - f(x))^2 \\
&= (g_f(x + \epsilon) + g_f(x) - 2f(x))(g_f(x + \epsilon) - g_f(x)) \\
&= (w^\top x + w^\top \epsilon + w^\top x - 2f(x))(w^\top x + w^\top \epsilon - w^\top x) \\
&= (2w^\top x - 2f(x) + w^\top \epsilon)w^\top \epsilon \\
&= 2(w^\top x - f(x))w^\top \epsilon + w^\top w \epsilon^\top \epsilon^\top, \\
E_\epsilon[F_{g_f} - F_0] &= 2(w^\top x - f(x))w^\top E_\epsilon[\epsilon] + w^\top w E_\epsilon[\epsilon^\top \epsilon], \\
\text{Since } \epsilon^{(d)} &\sim \mathcal{N}\left(0, (\sigma^{(d)})^2\right), \\
E[F_{g_f}] - F_0 &= E_\epsilon[F_{g_f} - F_0] = 0 + w^\top w \sigma^\top \sigma \geq 0.
\end{aligned}$$

□

Appendix B. Other Measures of Regularized Explanations



Appendix Figure B.1: Other measures of Regularized Explainer (Left) and Regularized Predictor (Right) in simulation study. X-axis is the regularization weight λ . Regularized Explainer does not affect model correctness over noise or model robustness while Regularized Predictor can improve them. Both regularized methods improve expected faithfulness and attribution stability under input uncertainty. Smaller value on y-axis is better.

Appendix Figure B.1 demonstrates how regularization weight λ influences measurements on model prediction and explanation, including model prediction correctness over noise, model prediction robustness, explanation expected faithfulness and attribution explanation stability.

Here we describe how the measures are defined for an instance, and the results in Appendix B are the average of these measures over all instances in the wine data set. The model correctness over noise of an instance is defined as the mean squared error between model prediction and ground truth over all hypothetical instances drawn from the input uncertainty distribution of the instance. Model prediction robustness of an instance is defined as the mean squared error between the model prediction of the instance and those of hypothetical instances drawn from the input noise distribution. Explanation expected faithfulness distance is the same as the definition in Section 4.3. Attribution explanation stability of an instance is the sum of the standard deviation of each feature attribution over the hypothetical instances. The fourth row in Appendix Figure B.1 is the sum of explanation stability over all features, and in the last row different features' attribution explanation stability is overlaid. All the error bars represent the standard error of the average over all instances. For all these measures, the smaller the values are the more correct, faithful, robust or stable the method is.

The left column in Appendix Figure B.1 shows that when increasing the regularization weight λ of the explainer, model correctness and robustness are not affected because LIME explainer is post-hoc and model-agnostic. The expected faithfulness distance of explanation increases with λ , and when λ becomes too big, the explainer's faithfulness distance gets worse. The stability of attribution keeps increasing with λ .

The right column illustrates that the model correctness of Regularized Predictor first improves with λ and then gets worse when λ becomes too big. Model robustness, explanation's expected faithfulness distance and stability keep increasing with λ .

Appendix C. User Interface of Survey

This appendix shows the various pages in the quantitative survey implemented in Qualtrics for Amazon Mechanical Turk workers as participants. Note that branching logic was used to manage the random assignment to different conditions and trial sequences. Each figure illustrates a page in the survey.

Rate wines with a Smart Wine Rating System!

If you were previously disqualified, feel free to try the HIT again.

If you have completed this HIT recently, thank you, but please do not accept it again. We may have to reject your newer work if we detect duplicate workers.

In this research study, you will use a Smart Wine Rating System to rate the quality of different wines based on their chemical properties and answer a few questions about your rating.

1. After giving consent,
2. You will go through a quick tutorial, then answer screening questions to check that you understand. You need to read carefully to correctly answer these questions to participate in the study.
 - If you pass the screening, you will answer questions to rate several wines and you need to answer demographic questions.
 - If you fail the screening, please return the HIT, since we cannot qualify you to continue with the HIT and do not want to give you a rejection.
3. You will test two versions of the System, where for each system
 - You will complete 15 quick decision making trials.
 - You can get an additional bonus of up to \$1.00 if you make at least 12 decisions correct within a 15-minute time limit.

Compensation

- Base compensation if you complete the survey after passing the screening questions is \$5.00.
- For both systems, you can get a total bonus of \$2.00.

The survey should take about 30–40 minutes.

Appendix Figure C.1: Welcome page introducing the tasks in the survey and compensation with bonus incentive.

Tutorial: Task Description

Imagine you are a quality control inspector at a winery. Your job is to:

1. Rate the wine quality (score: 0 to 100) of different wines based on its chemical properties and
 - **Accept** wines with **good quality** (score > 50, excluding 50) and
 - **Reject** wines with **poor quality** (score <= 50, including 50).
2. You need to inspect as many wines as you can, as quickly as possible. Aim to rate 30 wines in 30 minutes, i.e., inspect one wine in about 60 seconds or faster.
3. Avoid making mistakes, but it is OK to make some mistakes. Aim to be **at least 80% correct** in accepting or rejecting wines. That is, only 20/100 of your inspection decisions can be wrong.
4. You can get an **up to \$2.00 bonus** if you rate at least 80% wines correctly within the time limit. You will still get **the base compensation of \$5.00** regardless of your performance.

Screening Questions

You need to pass these test questions to qualify for the survey.

1. Suppose you rated the following wines with the indicated quality scores, which should you accept?
Hint: you may select more than one item.

80 30 55 42 50

2. If you rated 15 wines, what is the most mistakes you can make?

1 2 3 4 5

Tutorial: Wine Rating with Smart Wine Quality Rating System

The Smart Wine Quality Rating System automatically rates wine based on the chemical readings.

- You can use the system to help you to **quickly** accept or reject wines.
- You will only inspect wines that the system is **unsure** about, and therefore borderline cases that could have a score around 50.
- These uncertain wines tend to have **true** wine quality scores from **30 to 70**.
- Of these wines that you inspect, about half (50%) should be accepted, and half rejected.

Screening Questions

You need to pass these test questions to qualify for the survey.

3. If you *rated* (either accept or reject) 30 wines, about how many wines should you accept?

5 10 15 20 25

Appendix Figure C.2: First tutorial on task description and introduction to the AI system with screening questions.

Tutorial: Chemical Properties for Wine Quality Rating

The quality of a wine can be rated from **0 (very bad)** to **100 (very excellent)**. This relates to how a wine expert would rate the wine after tasting it. Instead of tasting, the Smart Wine Rating System measures the following chemical properties to estimate the wine quality.

Chemical Property	Generally, wines with ...	Reading	
		Min	Max
Alcohol (%)	Higher % alcohol has higher quality . But too much (>14) alcohol reduces quality.	8.3	14.9
pH (Basicity/Acidity)	Higher pH (less acidic) has slightly lower quality .	2.74	4.01
Sulphates (g/L)	More sulphates have higher quality , by controlling tartness and clarity. But too much (>0.85) slightly reduces quality.	0.33	2.00
SO ₂ Sulphites (mg/L)	More preservative sulphites have lower quality .	6	289
Vinegar Taint (g/L)	Higher vinegar taint has lower quality .	0.12	1.59

Note that these are *general trends*, but actual wine quality rated by experts may be different.

Screening Question

You need to pass these test questions to qualify for the survey.

Generally, which chemical properties increase wine quality?

Hint: you may select more than one item.

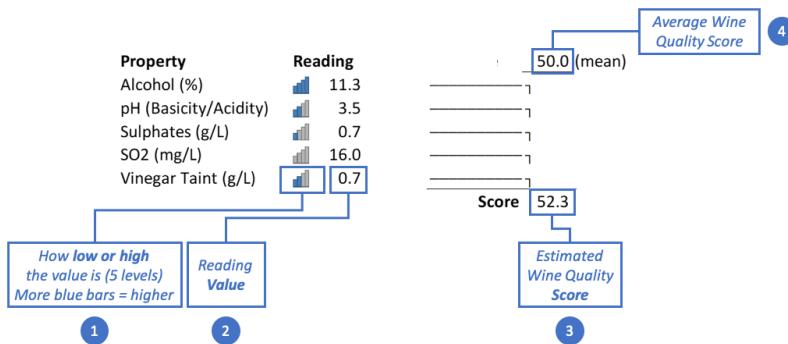
Alcohol pH Sulphates SO₂ Sulphites Vinegar Taint

Appendix Figure C.3: Tutorial on background knowledge on estimating wine quality rating based on chemical properties, with screening question.

Tutorial: Smart Wine Rating System (basic version)

The Smart Wine Rating System with chemical property readings and estimated wine quality score. We annotate the key features with blue text. It shows

1. Whether each reading is low or high (5 levels).
2. Each property reading value.
3. Estimated wine quality score. Note that it may be inaccurate.
4. Average wine quality score (this is always 50.0).



Screening Questions

Please study the user interface to answer the following questions.
You need to pass these test questions to qualify for the survey.

1. What is the reading for pH?
- 11.3 3.5 0.7 16.0 50 52.3

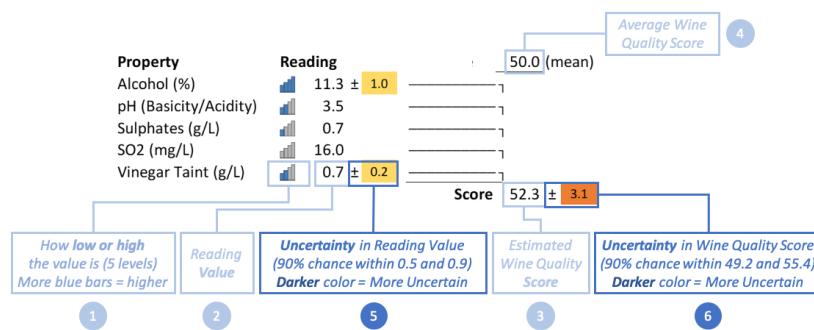
Appendix Figure C.4: First tutorial on AI system, regarding the basic version with no explanation (None), with screening question.

Tutorial: Smart Wine Rating System (with uncertainty)

Sometimes, the measurement of chemical properties can be inaccurate or uncertain.

Here is a version of the Smart Wine Rating System with chemical property readings with uncertainty and estimated wine quality score with uncertainty. We annotate the key features with blue text. It shows

1. Whether each reading is low or high (5 levels).
2. Each property reading value.
3. Estimated wine quality score. Note that it may be inaccurate.
4. Average wine quality score (this is always 50.0).
5. The uncertainty in the reading values. “ \pm ” indicates the range with a 90% chance that the actual value is within.
 - o For reading uncertainty, the color indicates how concerning the reading uncertainty is compared to the range of the chemical property (i.e. Alcohol range is 10.3 to 12.3, and Vinegar Taint range is 0.5 to 0.9 which is smaller), so the color may not match the number shown. **Use the background color to learn how big the reading uncertainty is.**
6. The uncertainty in the wine quality score.



Screening Questions

Please study the user interface to answer the following questions.

You need to pass these test questions to qualify for the survey.

1. What is the uncertainty of the wine score?

0.2 1.0 3.1 50.0 52.3

2. In the diagram above, **notice** that the background color of Reading Uncertainty is the same for Alcohol and Vinegar Taint. Are both readings equally uncertain?

Yes No

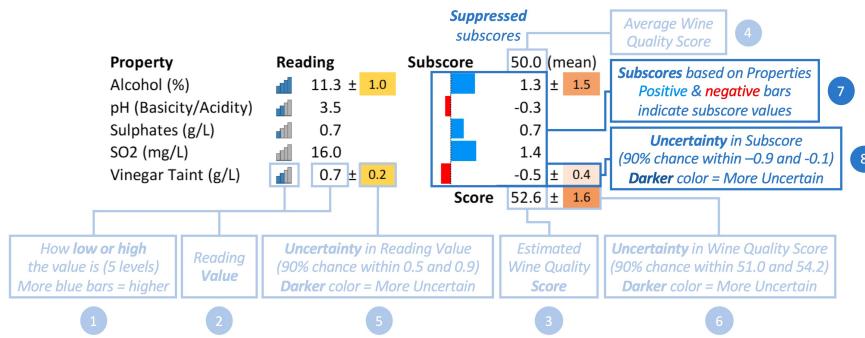
Appendix Figure C.5: Second tutorial on AI system, regarding the basic verison with no explanation (None), now also discussing uncertainty in input readings, with screening question.

Tutorial: Smart Wine Rating System (with suppressed subscore & uncertainty)

Here is a version of the Smart Wine Rating System with chemical property readings with uncertainty, **suppressed subscores**, and estimated wine quality scores with uncertainty. We annotate the key features with blue text. It shows

1. Whether each reading is low or high (5 levels).
2. Each property reading value.
3. Estimated wine quality score. Note that it may be inaccurate.
4. Average wine quality score (this is always 50.0).
5. Uncertainty in the reading values. "±" indicates the range with a 90% chance that the actual value is within.
 - o For reading uncertainty, the color indicates how concerning the reading uncertainty is, but this may not match the number shown. Pay attention to the color.
6. Uncertainty in the wine quality score.
7. Subscores indicating which property is estimated to **increase** or **decrease** the wine quality rating by some points. The mean and subscores add up to the total score.
 - o The system can also re-estimate the subscores by **suppressing** subscores if the reading uncertainty is high. In this case, both Alcohol and Vinegar Taint subscores are suppressed (made smaller with shorter bar). After suppressing, the system depends less on uncertain readings to estimate wine score.
8. Uncertainty in subscores, calculated based on the reading uncertainty. Suppression also reduces subscore uncertainty. The subscore uncertainties add as the sum of squares, e.g., $1.5 \times 1.5 + 0.4 \times 0.4 = 1.6 \times 1.6$.
 - o For subscore uncertainty, the color indicates how concerning the subscore uncertainty is. This matches the number shown.

Hint: Mouse over to see the unsuppressed version



Screening Questions

Please study the user interface to answer the following questions.

1. In the suppressed subscore version, which properties **decreased** the wine quality score?
Hint: you may select more than one item.

Alcohol pH Sulphates SO2 Vinegar Taint

Appendix Figure C.6: Pre-Condition tutorial on the system user interface showing relevant explanation components (ShowSuppress version shown with all components). Participants are incrementally introduced to different UI components to facilitate a gentle introduction. Questions are for manipulation check of comprehension, but not used for screening.

Congratulations! On to main survey

Congratulations! You have correctly completed the tutorial questions.
Please tell us a little about yourself.

Do you agree or disagree with the following statements?

	Strongly disagree	Disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Agree	Strongly agree
Uncertainty stops me from having a firm opinion.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The smallest doubt can stop me from acting.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I must get away from all uncertain situations.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
When I feel uncertain, I try to take decisive steps to clarify the situation.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
When I feel uncertain about something, I try to rationally weigh up all the information I have.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
When uncertain, I act very cautiously until I have more information about the situation.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Appendix Figure C.7: Post-screening qualification page with survey on Uncertainty Intolerance to measure confound of user personality background. This is asked before any experiment trials to avoid measuring a change in opinion due to performing potentially difficult tasks.

Sorry, you do not qualify to continue

Sorry, you have made mistakes in the tutorial questions. Please return the HIT to avoid getting a rejection. Thank you for trying!

Appendix Figure C.8: Post-screening disqualification page for participants who had 5 out of 7 correct answers to screening questions.

Version 1: Wine Rating 1

So far you have accepted 0 wines and rejected 0 wines.

000007

This wine has been measured and scored with the following information.

Readings and suppressed subscores:

Property	Reading	Suppressed Subscore	50.0 (mean)
			0.2 ± 1.0
Alcohol (%)	10.6 ± 0.9		
pH (Basicity/Acidity)	3.4	-0.4	
Sulphates (g/L)	0.6	-1.7	
SO ₂ (mg/L)	15.0	1.2	
Vinegar Taint (g/L)	0.7 ± 0.1	-0.5 ± 0.3	
Score		48.8 ± 1.1	

Hover over the image to see default subscores for comparison.

Note that this is a borderline case, and the system may be wrong, so you may not want to copy the displayed score.

The subscores + mean add up to give the total score.

The subscore uncertainties add as sum of squares, e.g. $2.8 \times 2.8 + 0.9 \times 0.9 = 3.1 \times 3.1$.

Please use the suppressed subscores version to make decisions.

As quickly and accurately as you can, please rate and decide whether to accept or reject this wine.

Your inspection **decision**

Reject (<=50) Accept (>50)

Your wine quality **score** (could be lower, equal to, or higher than system's score value)

30 32 34 36 38 40 42 44 46 48 50 52 54 56 58 60 62 64 66 68 70

Your wine quality **score uncertainty** (±) (could be lower, equal to, or higher than system's score uncertainty)

If you are **less sure** of your score, your uncertainty should be **larger** (not smaller)

If you are **more sure** of your score, your uncertainty should be **smaller**.

0 0.5 1 1.5 2 2.5 3 3.5 4 4.5 5 5.5 6 6.5 7 7.5 8 8.5 9 9.5 10

Here is the description of different chemical properties:

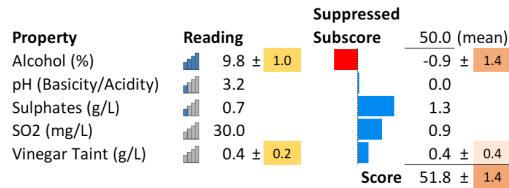
Chemical Property	Generally, wines with ...	Reading Min Max
Alcohol (%)	Higher % alcohol has higher quality . <i>But too much (>14) alcohol reduces quality.</i>	8.3 14.9
pH (Basicity/Acidity)	Higher pH (less acidic) has slightly lower quality .	2.74 4.01
Sulphates (g/L)	More sulphates have higher quality , by controlling tartness and clarity. <i>But too much (>0.85) slightly reduces quality.</i>	0.33 2.00
SO ₂ Sulphites (mg/L)	More preservative sulphites have lower quality .	6 289
Vinegar Taint (g/L)	Higher vinegar taint has lower quality .	0.12 1.59

Note that these are *general trends*, but actual wine quality rated by experts may be different.

Appendix Figure C.9: First page of a trial task (same for practice and main trials) showing timer, system user interface (UI), and three questions regarding inspection decision, score value, and score uncertainty. The UI is slightly different for different Explanation Technique conditions, showing the corresponding explanation components. The background table is included for reference.

Let us review your answers.

For this wine reading:



The correct wine rating is **64.5**, so the wine should be **Accept (>50)**.

You had scored 52.9 ± 2.1 .

Your decision to **accept** the wine was **correct**.

In a few sentences, please **reflect** on your reasoning for how you should make your decision. What did you look at and how did it affect your score estimation?

Appendix Figure C.10: Second page of a Practice trial task showing the participant's answers as a reminder and asking about their rationale in decision and rating estimation. This page was not included for the main trial. It serves two purposes: i) manipulation check to ensure that users are paying attention and reasonably understand the user interface, ii) stimulate users to self-explain to raise their level of understanding before proceeding to the main trials.

Version 1: Opinion about System Version

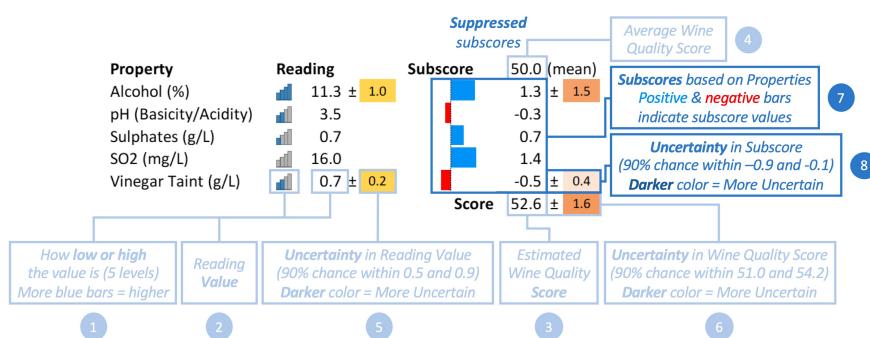
In the first system, you made correct decisions for **13/15** wines within the first 15 minutes (correct decisions after 15 minutes are not counted).

Please answer the following questions regarding the version of the Smart Wine Rating System that you used in the past 15 wines.

Do you agree or disagree with the following statements?

	Strongly disagree	Disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Agree	Strongly agree
I am confident of my decisions.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I trust the system's estimation of wine quality.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The system was helpful for me to make decisions.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Here is a reminder of the system user interface, showing one example wine.



Appendix Figure C.11: Post-Condition (after main trials) survey page with rating questions about the overall system.

Demographics

To complete this survey, please answer the following demographic questions.

Age

Gender

- Male
- Female
- Other (please specify)

What is your ethnicity?

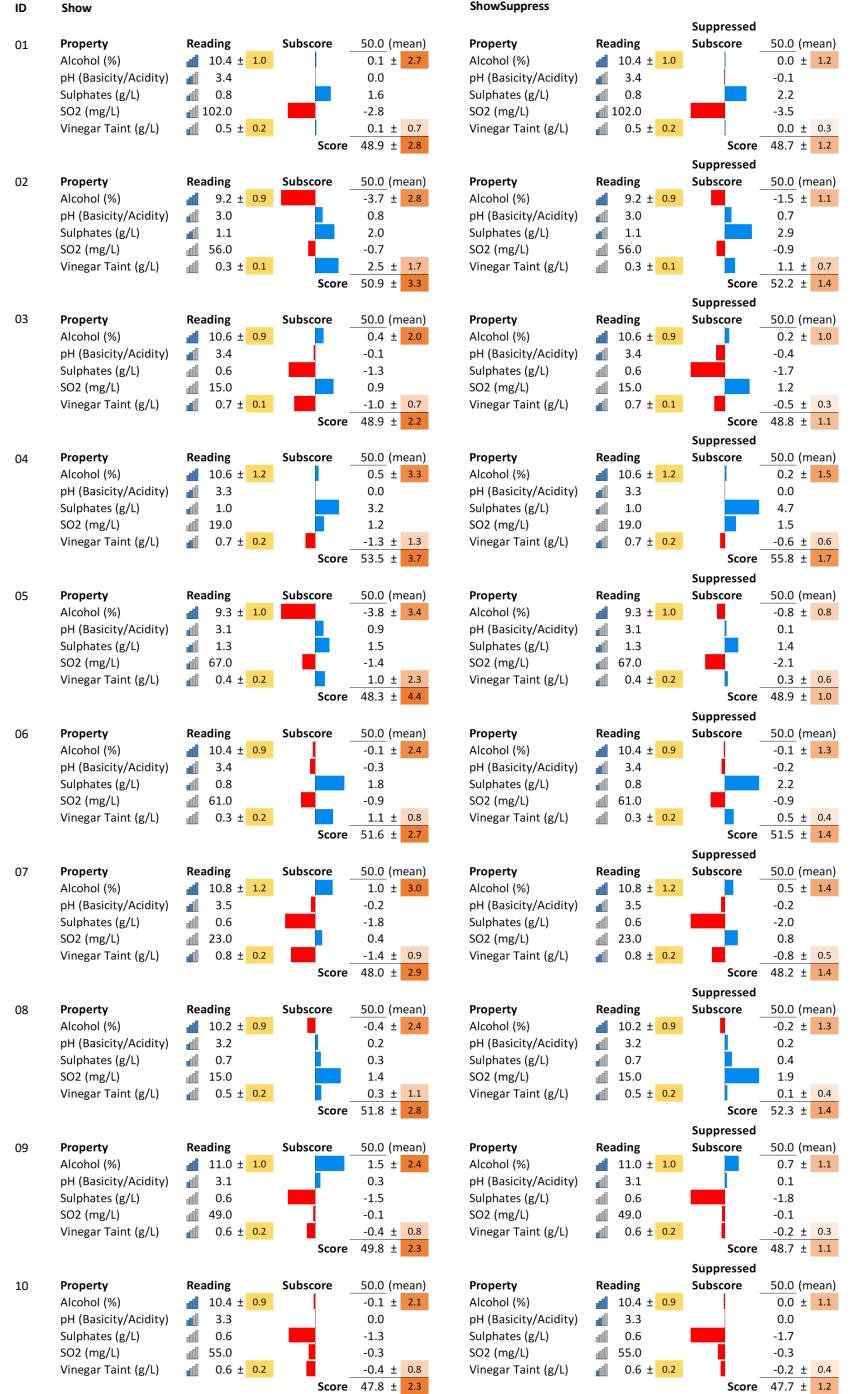
- White
- Hispanic or Latino
- Black or African American
- Native American or American Indian
- Asian or Pacific Islander
- Other (please specify)

What is the highest degree or level of education you have completed?

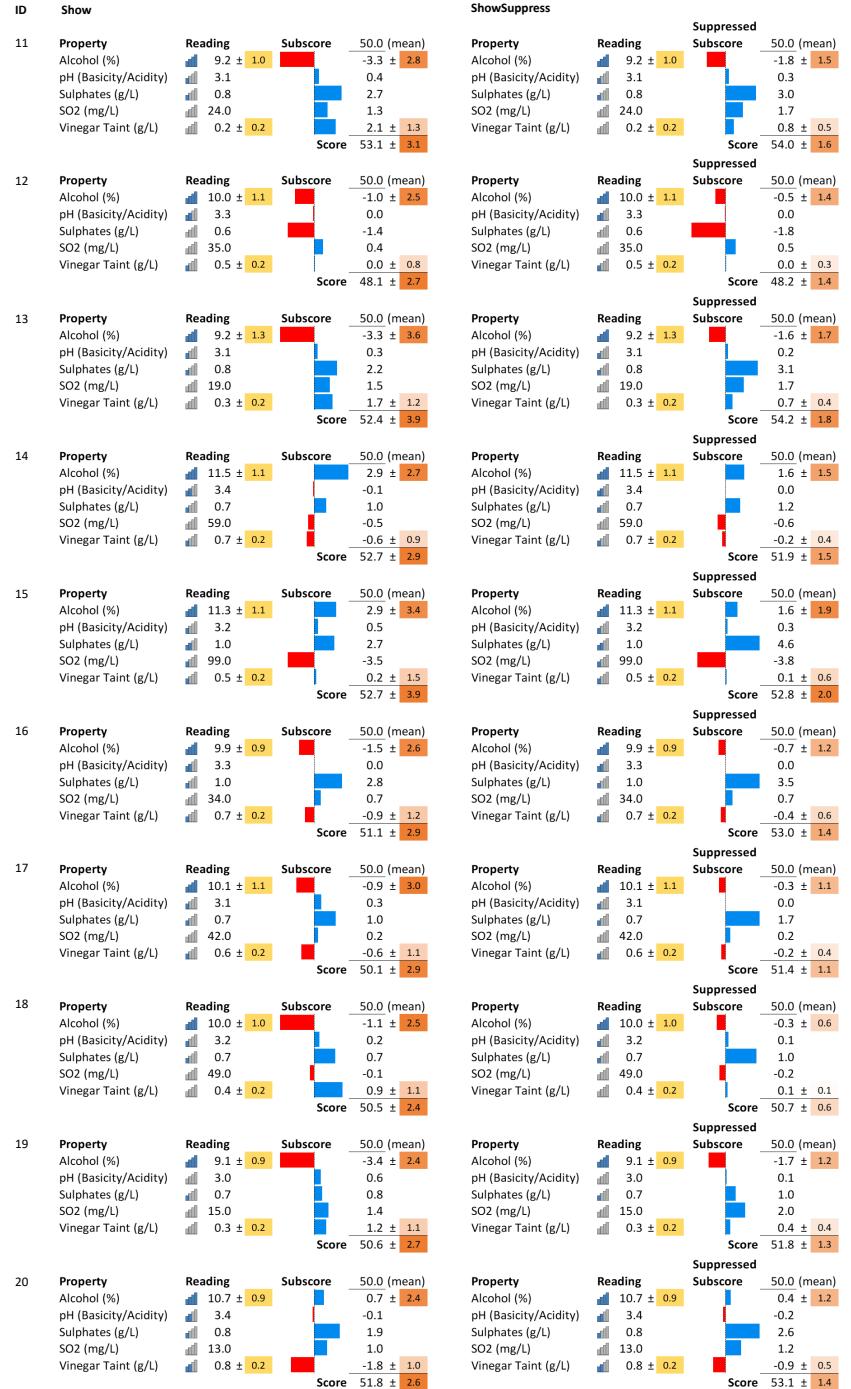
What is your current employment status?

Which of the following industries most closely matches the one in which you are employed?

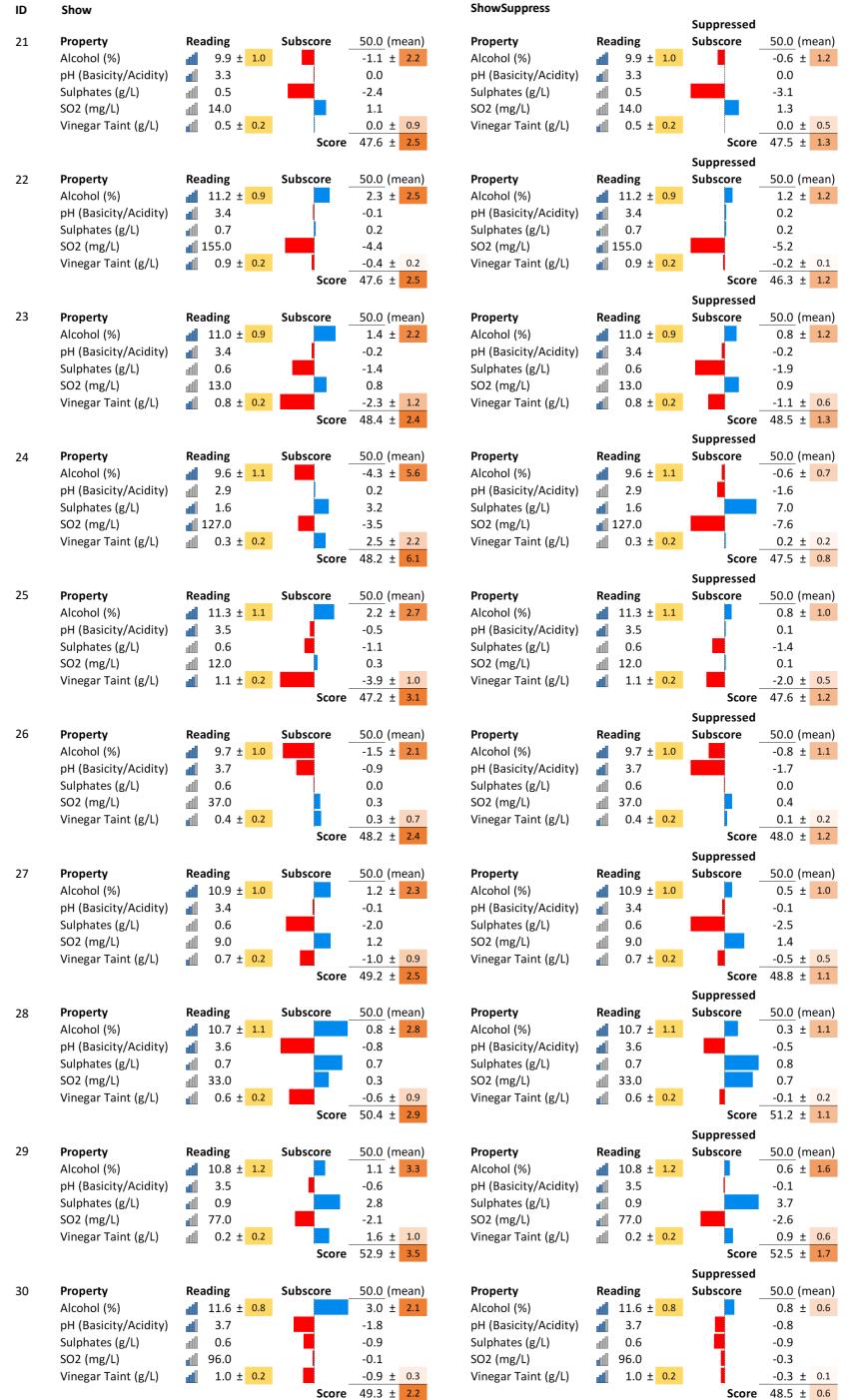
Appendix Figure C.12: Final survey page with questions on demographics.



Appendix Figure C.13: Instances used in user study. Here we show the interface of Show (left) and ShowSuppress (right) explanation.



Appendix Figure C.13: (continued) Instances used in user study. Here we show the interface of Show (left) and ShowSuppress (right) explanation.



Appendix Figure C.13: (continued) Instances used in user study. Here we show the interface of Show (left) and ShowSuppress (right) explanation.