Your Model Is Unfair, Are You Even Aware? Inverse Relationship Between Comprehension and Trust in Explainability Visualizations of Biased ML Models

Zhanna Kaufman, Madeline Endres, Cindy Xiong Bearfield, and Yuriy Brun

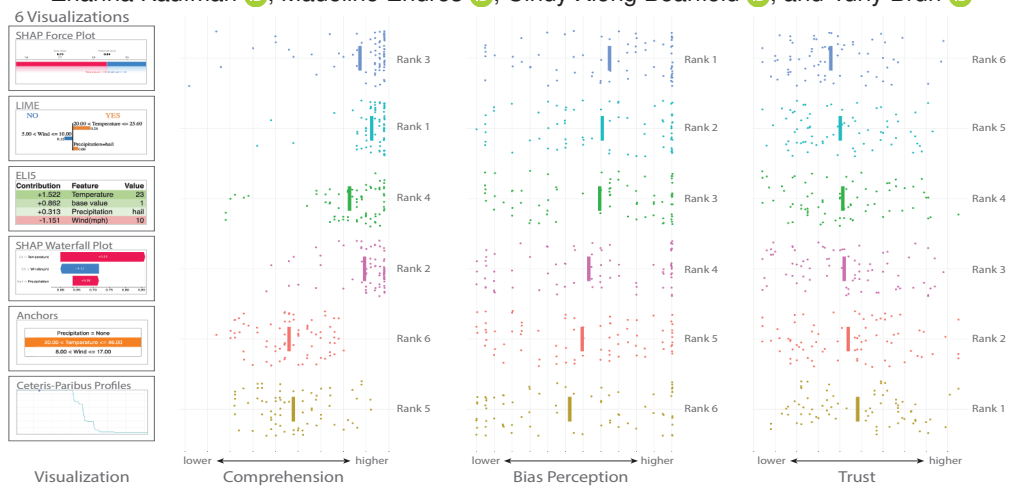


Fig. 1: The distributions and means of participants' comprehension, bias perception, and trust in ML models across five explainability visualization tools: SHAP, LIME, ELI5, Anchors, and Ceteris-Paribus Profiles. We find a negative correlation between comprehension and trust, strongly mediated by bias perception. That is, *the better* visualizations explain biased ML models, the less people trust those models. The strong negative correlation between bias perception and trust can be seen by comparing the middle and right graphs.

Abstract—Systems relying on ML have become ubiquitous, but so has biased behavior within them. Research shows that bias significantly affects stakeholders' trust in systems and how they use them. Further, stakeholders of different backgrounds view and trust the same systems differently. Thus, how ML models' behavior is explained plays a key role in comprehension and trust. We survey explainability visualizations, creating a taxonomy of design characteristics. We conduct user studies to evaluate five state-of-the-art visualization tools (LIME, SHAP, CP, Anchors, and ELI5) for model explainability, measuring how taxonomy characteristics affect comprehension, bias perception, and trust for non-expert ML users. Surprisingly, we find an inverse relationship between comprehension and trust: the better users understand the models, the less they trust them. We investigate the cause and find that this relationship is strongly mediated by bias perception: more comprehensible visualizations increase people's perception of bias, and increased bias perception reduces trust. We confirm this relationship is causal: Manipulating explainability visualizations to control comprehension, bias perception, and trust, we show that visualization design can significantly ($p < 0.001$) increase comprehension, increase perceived bias, and reduce trust. Conversely, reducing perceived model bias, either by improving model fairness or by adjusting visualization design, significantly increases trust even when comprehension remains high. Our work advances understanding of how comprehension affects trust and systematically investigates visualization's role in facilitating responsible ML applications.

Index Terms—Visualization design, explainability, trust, bias in machine learning

1 Introduction

Modern software systems increasingly rely on machine learning (ML), including in high-impact societal domains such as healthcare [52], hiring [80], banking [74], and criminal justice [4]. Stakeholders ranging from engineers and domain experts to policymakers and end-users routinely make critical decisions about these ML-driven systems [26, 40, 91]. These decisions span from choosing which ML technology to use, to how it is integrated, to which systems consumers trust and buy.

Unfortunately, modern ML systems frequently exhibit sexist, racist, and otherwise biased behavior [88]. For example, ML-based systems can have lower cancer detection rates for people of color [103], facial recognition systems can discriminate based on sex and race [15], and software often overestimates recidivism likelihoods for people of color [4]. Such biases have prompted legal restrictions on certain ML systems [83, 85]. Thus, to make informed decisions, stakeholders need to *understand* model behavior and detect potential *biases*. It is essential to help non-ML-experts accurately interpret and evaluate ML models.

Explainability visualizations, graphical representations clarifying ML model behavior, can help improve understanding [11, 17, 28, 38, 46, 63, 70]. For tabular data classifiers, local explainability visualizations highlight the importance of specific feature values (e.g., how age affects loan rejection). The left column of Figure 1 shows six such visualiza-

- *Zhanna Kaufman, Madeline Endres, and Yuriy Brun, University of Massachusetts. E-mail: {zhannakaufma, mendres, brun}@cs.umass.edu.*
- *Cindy Xiong Bearfield, Georgia Institute of Technology. E-mail: cxiong@gatech.edu.*

Received 31 March 2025; revised 1 July 2025; accepted 15 July 2025. Date of publication 5 December 2025; date of current version 3 February 2026. This article has supplementary downloadable material available at https://doi.org/10.1109/TVCG.2025.3634245, provided by the authors Digital Object Identifier no. 10.1109/TVCG.2025.3634245

tions: SHAP force [59] and waterfall plots [60], LIME bar charts [77], ELI5 tables [53], Anchors [78], and Ceteris-Paribus Profiles [12].

These visualizations facilitate user comprehension [8,77] and bias perception [31,90]. However, they are highly heterogeneous in how they indicate feature attribution (e.g., color, positioning, or shape), present alternative input instances, and communicate information (e.g., explicitly or implicitly). Design choices can profoundly affect reasoning [101], causal conclusions [100], and fairness perceptions [28,90,96]. Yet, there exists a limited systematic understanding of how visualization design influences comprehension, bias detection, and trust.

To address this need, we systematically explore how visualization design decisions affect stakeholders' ability to comprehend ML model behavior, accurately detect bias, and appropriately allocate trust. We focus on local explainability visualizations for tabular classifiers due to their widespread critical use (e.g., in healthcare [52], hiring [80], and finance [74]). Our goal is insights that generalize across heterogeneous models, capture perceptions from non-ML-expert users, and capture causality. These insights will enable the design of visualizations that align perception of bias with actual model behavior.

Our approach leverages two key insights. First, systematic study of state-of-the-art tools allows us to build a taxonomy capturing abstract design variability across visualizations, highlighting characteristics critical for comprehension and perception. Second, abstracted taxonomy features enable controlled evaluation to causally examine how visualization characteristics impact comprehension, bias detection, and trust, generating insights likely to generalize beyond individual tools.

We find that visualization design significantly influences comprehension, trust, and bias perception. Providing explicit feature attribution values significantly increases comprehension and bias detection. Simplifying visualizations (reducing the number of visual characteristics) also increases bias perception, even when comprehension remains constant.

Surprisingly, visualizations that increase comprehension, reduce trust. We investigate the cause of this relationship and find that it is strongly mediated by bias perception($p < 0.001$ for all): more comprehensible visualizations may increase bias perception, and a higher perception of bias reduces trust. We also demonstrate that specific design choices (e.g., implicit vs. explicit feature values) can enable more accurate detection of bias ($d = 0.22$, $p < 0.001$). Reducing actual model bias increases trust ($d = 0.44$, $p < 0.001$), but decreasing bias perception without altering actual bias and minimally impacting comprehension can still increase trust ($d = -0.19$, $p < 0.001$).

This paper makes the following contributions:

- A systematic review of 26 visualizations, producing a *comprehensive taxonomy of visualization design characteristics*, abstracted from state-of-the-art local explainability visualization tools.
- A series of *large-scale user studies* with 818 non-expert participants, empirically assessing how design characteristics correlate with and causally influence comprehension, bias perception, and trust.

Our stimuli and data are available in our supplementary materials [47].

2 Background: ML Explanation Visualizations

As ML use has become more common, the importance of non-experts understanding ML model behavior has increased. Non-experts often need to understand and use ML models [6,81,106] and non-experts' input can improve ML automation [20,30,67], creating a need for effective explainability tools [11] that target non-experts [11,38,63,70].

The needs for ML explainability are significant. End-users often have difficulty understanding ML decisions [36], which can deteriorate their trust and use of ML tools [22]. The European Union's General Data Protection Regulation even requires subjects of decisions made by ML systems to have a right to an explanation [25]. Additionally, system designers need to understand ML models when incorporating them into their systems. In practice, however, many ML models, including deep learning, are so complex that even ML experts struggle to understand their functionality [82]. Debugging unexpected behavior, particularly when that behavior is driven by complex, opaque ML models, can be labor-intensive and slow [58]. Biased model behavior, which is unfortunately common in ML [29,88,92] adds an extra layer of complexity; bias can be particularly difficult to formalize and control during training because of a possible trade-off between fairness and accuracy [56], and different notions of bias can be incompatible [27].

Visualization offers hope. Visualizing model training can support debugging [58], and improve explainability for hard-to-formalize properties, such as safety [16]. It can also clarify behavior when training and deployment environments differ [57], and support the detection and debugging of bias [2,45,46]. However, few insights exist into how visualization design choices influence comprehension, bias perception, and trust. Providing these insights is the goal of this paper.

2.1 ML Explainability

This paper focuses on the most common type of ML models: classification models for tabular data. We use the term *model* to refer to classification models whose goal is classifying inputs into categories, e.g., images of medical tissue scans into cancerous or non-cancerous. More formally, an ML model is a mathematical function mapping feature sets to categorical labels. ML explainability methods generally fall into three categories: inherently interpretable, global, and local.

Inherently interpretable (white-box) models are self-explanatory by design [7], but often fail to capture intricate feature interactions. By contrast, many high-performing models are black-box, requiring model-agnostic approaches to explain behavior [9,14,34,60,77]. *Global explanations* use approximations to summarize model behavior across all predictions [7,34,42,60,65,77,86,93], offering useful high-level insights. However, they often lack the granularity needed to assess fairness or trust in individual predictions.

Local explanations describe model behavior for individual input instances [60,77,86]. There are many local explanation types, which vary substantially in visual design and emphasis. Feature importance scores quantify how each feature impacts the model's predicted probabilities [93]. Ceteris-Paribus (CP) profiles, for each input, graph changes in predicted model output (y axis) as a result of changing a single feature in isolation (x axis) [13]. Individual conditional expectation (ICE) plots combine multiple CP profiles for the same feature, each representing a different input [32]. Contrastive explanations compare model outputs when feature values change [87]. Counterfactual explanations identify feature-value modifications that alter predictions [33,97]. Finally, pertinent negatives find the minimal feature changes needed to flip model predictions, while pertinent positives show the smallest feature subset that must remain fixed to maintain the same prediction [51].

This paper investigates how local explanation visualizations facilitate comprehension and trust among non-expert ML users, particularly when models are unfair. We focus on local explanations because they provide granular logic for model behavior: a global explanation may highlight important features, but local explanations reveal how individual feature values shift model predictions. Local explanations require the user to infer broader patterns from individual examples [98], making them especially relevant for understanding bias perception and trust. We focus on tabular data classifiers because of their significant design heterogeneity. This provides a rich opportunity to examine how design influences comprehension and trust, particularly for unfair models. Specialized explanations for image and text classifiers are often visually similar across tools [93] (e.g., highlighting salient words or pixels).

While several explainability tools focus specifically on revealing bias during or after model development [2,17,31,44,46,95,99], they target programmers and data scientists. They require ML and bias expertise, include complex interfaces, and require prior understanding of model functionality. By contrast, we target a broader range of users, including non-experts. We thus focus on simpler tools that explain model behavior without requiring specialized background knowledge.

2.2 Explainability, Comprehension, and Trust

We now overview current understanding of user perceptions of visualization explanations, focusing on model comprehension and trust.

Explainability and Comprehension. Assessing comprehension is essential for understanding explainability visualization effectiveness because comprehension is a measure of model interpretability. However, comprehension remains surprisingly understudied; only 22% of

AI explainability tools include user studies measuring model comprehension [68]. Common methods include (1) asking participants to compare features' predictive power [19], predict model outputs [18,65], or match a prediction to an explanation [50], (2) measuring how often users made correct decisions when advised by a model [3,41,84] or detecting incorrect predictions [50], and (3) using open-ended questions about model decisions [77]. Generally, these studies suggest that explanations improve comprehension. However, due to the wide variety of tool designs and evaluation methods, comparing comprehension across different explanation visualizations remains challenging.

In this work, we integrate these approaches into a metric that allows comprehension comparisons across explainability visualizations. We define comprehension as an understanding of the model's output for a given input, along with the magnitude and direction of each input feature towards a specific classification outcome.

Explainability and Trust. The most common measure used for ML model *trust*—willingness to accept a model's outputs—often does not reflect people's actual trust levels [1], particularly when unfairness is involved. People may accept a model's recommendation, but distrust it due to awareness of broader socio-organizational contexts [23]. Furthermore, users may prefer AI recommendations over human ones, even when perceiving the AI less morally trustworthy [89]. By contrast, we define trust as a combination of perceived accuracy and willingness to rely on a model for decisions affecting users and others.

Prior studies report mixed findings on interpretability's effect on trust. Decision explanations can increase trust in model accuracy [77,105], and bias-focused explainability tools may foster trust [31]. In some domains, the mere presence of explanations can make people more likely to use an ML model [8], while in others, experts use their prior knowledge to calibrate their trust in a model based on its explanation [94]. Conversely, observed accuracy can have a more significant effect than interpretability [1] on both user trust and willingness to use decision-making assistants [76,76,107,108]. Increased comprehension does not necessarily promote model use [73]; practitioners can misunderstand and over-trust visualizations, mistakenly dismissing suspicious results [48], leading to misguided decisions [37,61]. Finally, perceptions of explainability methods can vary by audience demographics, affecting trust and bias perception [5,28,55].

Visualization design shapes both people's perception of data and their decision-making strategies [96,100,101]. Different visualization types affect trust, with familiar, simpler, or aesthetically pleasing visualizations increasing trust the most [21,24]. Visualization design can impact emotional reactions, driving important decisions [104]. It can also reveal model biases [90] and affect reactions to biased models [28]. This paper fills the need for further research into how specific design characteristics affect the relationship between comprehension, bias perception, and trust. We perform this investigation using the six state-of-the-art explainability visualizations seen in Figure 4.

3 Research Questions

We organize our investigation of how explainability visualizations can impact comprehension and trust around three research questions:

RQ1: What design characteristics of ML explainability visualizations could impact comprehension, trust, and bias perception?

RQ2: Do visualization design characteristic variations correlate with differences in model comprehension, bias perception, and trust?

RQ3: Are the observed relationships between comprehension, bias perception, and trust causal? Do the relationships generalize beyond existing tools?

We address RQ1 via the construction of a visualization characteristic taxonomy (Section 4). We then design and conduct user studies (Section 5) to address RQ2 (Section 6) and RQ3 (Section 7).

4 RQ1: What Visualization Characteristics Exist?

We want to identify visualization design characteristics that could impact model comprehension and trust. We desire an understanding that (1) covers state-of-the-art tools and (2) abstracts key characteristics that can be varied experimentally. We systematically analyze existing explanation visualization tools to construct a comprehensive taxonomy.

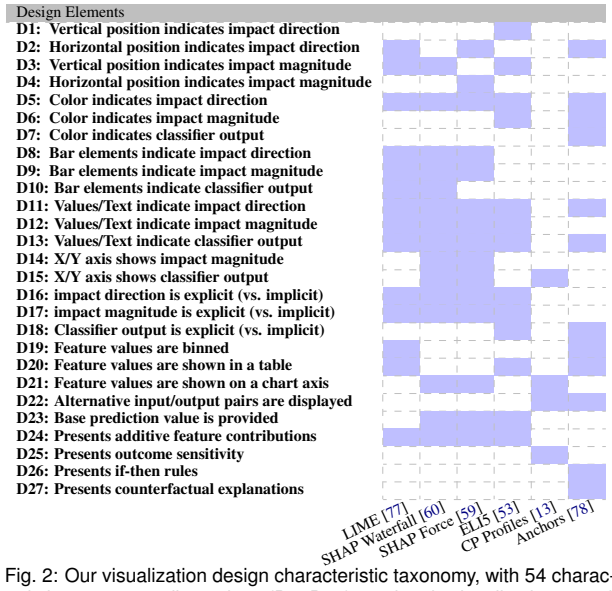


Fig. 2: Our visualization design characteristic taxonomy, with 54 characteristics across 27 dimensions (D1–D27), on the six visualizations used in our user study (Section 5.2.1). Shaded cells indicate that a specific dimension was characterized as true, while unshaded cells indicate it was characterized as false. Our supplementary materials [47] include the complete taxonomy applied to all 26 analyzed visualizations.

4.1 Taxonomy Development Methodology

To build our taxonomy, we first compile a representative collection of state-of-the-art local explanation visualization tools. We start with four recent surveys of visualization tools [16,62,68,79] (including one from 2024), supplemented by an updated list of explainability resources [35]. From these sources, we select all tools with publicly available implementations (enabling practical evaluation), excluding tools that provide purely textual explanations, non-local explanations, and explanations for non-tabular data. Through this method, we collect a set of 26 visualizations from 20 visualization-based ML explainability tools, which we empirically analyze using the taxonomy development methodology described by Nickerson et al. [69].

Nickerson et al. [69] define a taxonomy as a set of n dimensions, each consisting of k mutually exclusive characteristics, such that every classified member has exactly one characteristic per dimension. Following their empirical approach, we begin by defining our *meta-characteristic*, a central purpose from which all other characteristics derive. Using our comprehension definition from Section 2.2, we derive our meta-characteristic: *How does a visualization convey information about the impact of each individual feature on a single model prediction?*

We follow Nickerson et al.'s suggested iterative methodology. The first author analyzes progressively larger subsets of visualizations to identify characteristics related to our meta-characteristic (e.g., color, position, graph, etc.), consulting with co-authors at intermediate steps to establish consensus. These characteristics are then grouped into dimensions, such that each visualization has exactly one characteristic per dimension. This iterative process continues until the dimensions comprehensively reflect our meta-characteristic.

4.2 A Taxonomy of Visualization Design Characteristics

Figure 2 shows the 54 characteristics we found, grouped into 27 dimensions, D1–D27, with 2 characteristics per dimension (true/false). This figure also shows our encoding for the explainability visualizations in Figure 4. The dimensions are summarized here:

Metric	Type	No.	Text
Comprehension	Multiple Choice	C1	Will this model approve the loan for this person?
		C2	What feature has the most predictive power for this decision?
		C3	Which factor(s) are pushing the model toward predicting 'NO'/'YES'
Perceived Comprehension	Likert	PC1	How well did you understand the way this model makes decisions?
		PC2	How easy was it for you to understand the model output?
		PC3	How likely would you use this visualization to explain models to other people?
Trust	Likert		On a scale of 1 to 6, how much do you trust the model to approve or deny a loan …
		T1	… for you?
		T2	… for other people in general?
		T3	I trust the data this model was trained on.
		T4	This model is accurate.
		T5	Computer models can be trusted to make human decisions.
Bias Perception	Yes/No	B1	Do you think this model includes potentially discriminating factors?
		B2	This model uses all of the features that it should use when making this decision.
		B3	This model does not use any unnecessary features when making this decision.
		B4	This model is fair.
Behavioral Alignment	Yes/No		This model would probably give me a loan …
		A1	… because I am similar to the person described in this question.
		A2	… because I am different from the person described in this question.
		A3	… because of previous decisions it has made.
		A4	This model would probably not give me a loan, and this would be the correct decision.

Fig. 3: Overview of survey questions for our three metrics—Comprehension, Trust, Bias Perception—along with questions assessing if participants perceive that model behavior aligns with their expectations.

- Visualization element *colors* convey the direction **(D5)** or magnitude **(D6)** of feature impact, or the value of the prediction **(D7)**.
- *Bar elements* in the visualization convey the direction **(D8)** or magnitude **(D9)** of feature impact, or the value of the prediction **(D10)**.
- *Printed numerical values* show direction **(D11)** or magnitude **(D12)** of feature impact, or prediction values **(D13)**.
- Magnitude of feature impact **(D14)** or the value of the prediction **(D15)**, are on a numerical *x or y axis*.
- Feature impact direction **(D16)** and magnitude **(D17)**, or prediction values **(D18)** are conveyed *explicitly* or *implicitly*.
- Feature impacts are specific to individual values or generalized within value *bins* **(D19)**.
- Individual input *feature values are located* in a separate table, or indicated on the axis of a graph element **(D20–D21)**.
- Visualization may provide *alternative example feature inputs* and resulting predictions **(D22)**.
- Visualization may provide a *bias/base value* for predictions **(D23)**.
- Visualizations present different explanation types, including *Additive feature contributions, outcome sensitivity to input changes, counterfactual explanations*, and *if-then rules* **(D24–D27)**.

Our taxonomy of visualization design characteristics enables both evaluating existing explainability visualizations and designing new methods of conveying this information. For example in RQ3, we show how characteristics such as explicit indications of impact magnitude and classifier output (D17 and D18) can impact user comprehension and bias perception of the underlying model.

RQ1 Summary: Explainability Visualization Characteristics

Our taxonomy analysis identified 44 characteristics of visualization design characteristics across 26 ML explainability visualizations from 20 visualization tools intended to promote comprehension of tabular data classifiers (Figure 2). These characteristics allow us to systematically assess the impact of design decisions on user comprehension and perception of explainability visualizations.

5 RQ2 and RQ3 User Study Design

We next investigate how visualization design elements influence model comprehension, bias perception, and trust. To do so we conducted a series of online Qualtrics [75] surveys with crowd-sourced participants from Prolific.com [71]. All our studies were ethics-board approved.

Each survey follows the same format; participants are shown a series of explainability visualizations for a classifier that recommends whether to give a loan applicant a loan based on various demographic features (e.g., age, education, sex, etc.). For each loan applicant, participants respond to multiple-choice questions designed to assess model comprehension, bias perception, and trust. The surveys differ in either the explanation visualization used or the underlying model and its fairness.

In total, we conducted eleven surveys: six for RQ2 and five for RQ3. For RQ2, since we are interested in understanding if visualization design characteristics across state-of-the art tools correlate with model comprehension, bias perception, and trust, we present participants with explainability visualizations generated by one of six state-of-the-art tools, carefully chosen due to their high coverage of taxonomic characteristics (see Section 5.2.1).

For RQ3, as we wish to understand if the observed relationships are causal and if the impact of design characteristics on comprehension, bias perception, and trust generalize beyond existing tools, we use five survey variations with altered explainability visualizations to conduct *three crowd-sourced follow-up experiments designed for controlled* analysis of causality (see Section 7):

- *Experiment 1—Explicitness:* Increase the explicitness of an implicit visualization to modulate comprehension and confirm a causal effect on bias perception.
- *Experiment 2—Fairness:* Increase fairness of the underlying model to modulate bias perception and confirm a causal effect on trust.
- *Experiment 3—Bias Perception:* Manipulate design characteristics to alter bias perception while keeping comprehension high to confirm bias perception as the mediating factor.

Section 5.1 details our metrics for measuring comprehension, bias perception, and trust; Section 5.2 describes our explainability visualization stimuli; and Section 5.3 summarizes our participants.

5.1 Experimental Measurements

Using a series of multiple-choice questions after each explanation visualization scenario, we measured three primary aspects: comprehension of the underlying ML model, perceived bias in the model, and subsequent trust of that model. We also assessed behavioral alignment to investigate if people's perception of model behavior correlated with their trust. We developed our metrics by reviewing measurement methodologies in existing literature. In addition, we ran two pilot studies, each with over 200 participants, to assess our preliminary questions and finalize our metrics. We now discuss each metric in detail.

Measuring Comprehension: We use the definition of comprehension described in Section 2.2, operationalized via questions C1–C3 in Figure 3. We define an aggregate comprehension score as the sum of correct responses to these questions across all prediction instances. We further measure *perceived comprehension* using three Likert-style questions (PC1–PC3) which assess if people's perception of explainability visualization effectiveness matches their observed comprehension level.

Measuring Trust: We define trust in Section 2.2 as a person's perception of the accuracy of an underlying model and their willingness to rely on the model for decisions that affect themselves and others. We operationalize this definition via the questions T1–T5 in Figure 3, which participants answer for every visualization instance. We define the aggregate trust score as the sum across all 7 prediction instances.

Measuring Bias Perception: Bias perception measurement tends to be either qualitative or self-reported (e.g., asking participants to identify systemic unfairness [66], rate their general perception of fairness [66], or determine if certain subgroups are treated unfairly [19, 95]). We operationalize perceived bias as a positive response to B1 in Figure 3, combined with the sum of disagreement with statements B2–B4.

Behavioral Alignment: People can trust a model more when its decisions benefit them [28]. We summarize this potential behavioral alignment with the Yes/No questions A1–A3 in Figure 3. We also add A4, allowing participants to indicate that the model's decisions would not be beneficial to them, but they still approve of model behavior.

Qualitative Analysis: We asked participants two free response questions regarding which visualization characteristics they found most and least useful when answering questions about model behavior and model fairness. This allows us additional insights into why certain design characteristics may be associated with higher comprehension or trust.

Statistical Methods We conduct our analysis in RStudio [72]. To assess relationships between aggregate scores across visualizations in RQ2 we use linear regression. We use Analysis of Variance (Anova) to determine whether our independent variables are significant predictors of our outcomes, and estimated marginal means (emmeans) to compare average scores for our measures across visualizations. To compare measures across pairs of surveys in RQ3, because our distributions are not normal, we use Wilcoxon Rank Sum Tests. We consider the results significant if $p < 0.05$. For effect size, we use Cohen's d. Our full analysis can be found in our supplementary materials [47].

5.2 Explainability Visualization Stimuli

Each survey presented participants with 7 scenarios featuring a loan-recommending model. Each scenario consisted of an input (a person wanting a loan), the model's recommendation (approve or deny the loan), and an explanation visualization. The model was a LightGBM classifier [64] trained on the Census Income dataset (14 demographic features and income for 48,842 people) [10]. LightGBM classifiers are black-box models that use a gradient boosting decision tree algorithm [49]. To reduce visualization complexity, we used a subset of 5 features: age, education level, occupation, hours worked per week, and sex[1]). The model predicts income given these features. If actual income exceeds the prediction, the model recommends granting the loan. Because this dataset gives loans to 31% of men but only 11% of women, training on it without fairness constraints results in a model that is more likely to recommend a loan to men than to women.

We therefore focused on sex as the model's discriminatory factor, and included 3 visualizations for males and 3 for females, as well as a juxtaposition visualization for a male and a female with equivalent values for all features except sex. We chose the applicants (all actual data points from the Census Income dataset) to include a range of education, occupations, and hours worked per week. For the two juxtaposed applicants, we filtered the dataset for instances identical in every feature except sex, but for which the model produced different recommendations (one applicant received a loan and the other did not).

While each survey contained the same seven scenarios, the visualization differed. For RQ2, we consider six state-of-the art explanation visualizations with high coverage of our taxonomy, For RQ3, we consider five additional visualization variations, centered around three experiments testing for causality (see Section 5). We detail the RQ2 visualizations in Section 5.2.1) and the RQ3 variations in Section 5.2.2.

5.2.1 State-of-the-Art Explainability Visualizations (for RQ2)

In RQ2, we investigate how design differences across state-of-the art local explanation visualizations correlate with comprehension, bias perception, and trust. While we used 26 visualizations to build our taxonomy (see Figure 2), we selected six visualizations from popular tools that span the vast majority of taxonomy design characteristics: SHAP waterfall plots [60], SHAP force plots [59], ELI5 tables [53], LIME visualizations [77], Ceteris-Paribus (CP) profile plots [7], and Anchors explanations [78]. We do so to avoid redundancy and diminishing returns, as many visualizations share overlapping characteristics

For each visualization, we conduct a survey assessing comprehension, bias perception, and trust for the same seven loan-recommendation scenarios. To ensure ecological validity, we did our best to keep the visualizations as similar as possible to those produced by each state-of-the-art tool. However, due to the multiple configuration parameters available, combined with heterogeneous approaches to computing and ordering feature importance [54], we had to standardize non-visualization-related differences so that we can make conclusions regarding the effect of visual design. For example, we ensured that all visualizations had the same feature importance ordering and contribution values for a given scenario (using the SHAP order as our default). We also made minor standardization changes to facilitate understanding of the impact of visual design, rather than textual differences such as terminology (e.g., "base value" vs. "bias value"), or capitalization. All of our standardizations are in our supplementary materials [47].

[1]We use the term sex and the categories male and female, as in the dataset.

Figure 4 shows an example of each explanation visualization, as included in our surveys. We briefly describe each, highlighting key taxonomy characteristics motivating inclusion in our user study:

SHAP waterfall plots, Figure 4a: SHAP value attribution shows each feature's conditional contribution to the model's output, summing to the final predicted probability. SHAP is based off of shapley values in game theory, which are the solution to the equation $p(\text{output}) = b + \phi_1(\text{feature}_1) + \phi_2(\text{feature}_2) + \phi_3(\text{feature}_3) + \cdots$, where each ϕ is a shapley value. In the waterfall plot [60], colored arrow bars indicate positive (red) and negative (blue) feature contributions, stacked bottom-to-top from least to most significant. Starting at a base value, contributions shift the cumulative sum right (positive) or left (negative), with the top bar showing the final predicted probability. SHAP waterfall is one of only two visualizations in our study where bar elements indicate classifier output.

SHAP force plots, Figure 4b: These show the same information as waterfall plots, but as a single bar where red (positive) and blue (negative) sections represent feature contributions [59]. Sections are ordered around the predicted value, with positive contributions left and negative right. More significant features are closer to the center. SHAP force is the only visualization in our study where horizontal position indicates feature impact magnitude.

ELI5, Figure 4c: A Python package for explaining classifier predictions, ELI5 shows feature contributions as a table with colored rows: large positive contributions at the top (deep green) and large negative ones at the bottom (deep red). Each feature has a contribution score, computed in a model-specific manner. For LightGBM, ELI5 traces ensemble decision tree paths to determine how each step impacts the prediction. Contributions sum to an output score, though this score does not directly represent the model's predicted probability [53]. ELI5 is the only visualization in our study where vertical position indicates feature impact direction, and one of only two where classifier output is explicit or where color indicates feature impact magnitude.

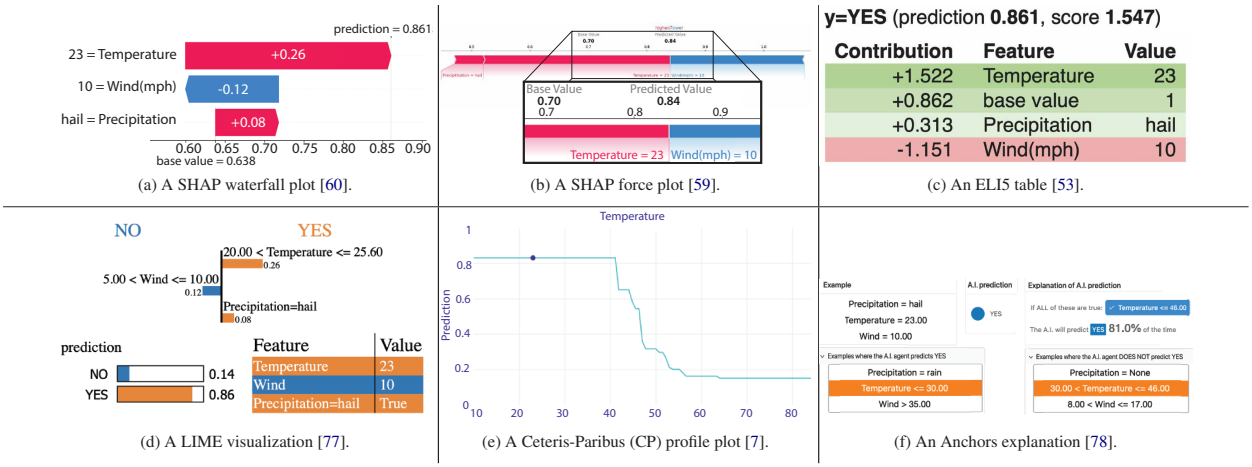
LIME, Figure 4d: LIME explains model predictions by (1 generating "near-by" input-output pairs by input mutation, and (2) fitting a white-box linear regression to estimate feature weights [77]. It visualizes feature weight magnitude and direction as a bidirectional horizontal bar chart, with features ordered by impact. Positive contributions are orange and point left; negative ones are blue and point right. LIME also includes a bar chart with class prediction probabilities, and a table where row color and placement show feature significance. LIME is one of only two visualizations in our study where feature values are binned or bar elements indicate classifier output.

CP profiles, Figure 4e: In a separate plot for each feature, CP profiles show how changing a feature affects model output, while holding other features constant (line graphs for continuous features, bar charts for categorical ones) [13]. While not explicit, average output change, measured via CP oscillation, reflects feature impact [13]. Our tutorial explained how to estimate this, with larger oscillations indicating stronger impact. We use the Python Dalex implementation [7], where dark blue marks the current input-output and teal lines or bars show how output varies by input. CP is one of only two visualizations in our study where feature impact is inferred, and the only one to not use color or text to indicate feature impact direction.

Anchors, Figure 4f: Anchors explain predictions by identifying feature constraints that, when satisfied, cause the model to consistently output the same prediction [78]. Each anchor is a rule specifying the constrained features and its associated precision. Explanations are primarily textual, with color (orange or blue) highlighting constrained features. It also includes example inputs both satisfying or violating the anchor, showing the effect on model output. Anchors is one of only two visualizations in our study where feature values are binned, and the only one where color indicates classifier outcome.

5.2.2 Adjusted Explainability Visualizations (for RQ3)

While our analysis for RQ2 can establish correlations between comprehension, bias perception, and trust that are grounded in taxonomy design characteristics, it cannot establish causality. Establishing causality is important to assess if findings are likely to generalize beyond



(a) A SHAP waterfall plot [60]. (b) A SHAP force plot [59]. (c) An ELI5 table [53].

(d) A LIME visualization [77]. (e) A Ceteris-Paribus (CP) profile plot [7]. (f) An Anchors explanation [78].

Fig. 4: The six state-of-the-art explainability visualizations in RQ2's user study: (a) SHAP waterfall, (b) SHAP force, (c) ELI5, (d) LIME, (e) CP, and (f) Anchors. We conducted six surveys on the same seven loan-recommendation scenarios, one with each visualization. Here, each explains the same input for a ML model that predicts if one should wear a coat based on the weather. We used this model in a tutorial at the start of our surveys.

specific existing explanation visualizations, thus informing the design of next-generation visualization designs.

In RQ3, we consider three causal experiments, one regarding the importance of explicit values, one assessing the impact of model fairness, and one on the relationship between bias perception and trust. We describe each experiment in detail in Section 7. However, at a high level, each involves two visualization surveys that are identical, other than a single visualization characteristic or model property, admitting causal comparison. Figure 8 shows examples of each pair of visualizations. While six surveys are analyzed in RQ3, we only run five additional surveys; since one condition for our explicitness experiment is a non-modified CP profile, we reuse its survey results from RQ2.

5.3 Recruitment and Population Contextualization

We recruited 825 participants from the crowdsourcing platform Prolific.com [71] who met our inclusion criteria (fluent in English, at least 18), 75 per survey. After additional quality filtering beyond Prolific's guarantees (e.g., attention checks, skipping questions), we had 818 valid participants (age 18–77); 438 for RQ2, and 380 for RQ3.

Our sample size was guided by power analyses on preliminary data (n=50). Using balanced one-way analyses (power 0.9, effect size from pilot data, assumed normality), we estimated that 35–68 participants per visualization would be sufficient to detect effects for individual survey questions. While our final analysis used composite scores instead, this approach gave us confidence that we would have enough statistical power to convincingly answer our research questions.

We used Prolific's interface to recruit equal numbers of participants identifying as men and women, as prior work suggests that gender may influence bias perception and trust [28]. 397 identified as cis or trans women, 395 as cis or trans men, 19 as non-binary, 2 as an unlisted gender, and one as unsure/questioning. Participants varied in education, income, and ethnicity. Regarding ML familiarity, participants were primarily non-experts: 219 had none, 389 were beginners, 189 had intermediate knowledge, and 22 were experts. Participants were compensated $12.00 an hour, consistent with Prolific recommendations.

6 RQ2: Does visualization design correlate with comprehension, bias perception, and trust?

Having defined key characteristics of local explanation visualizations (see Section 4), we now examine if these characteristics correlate with model comprehension, bias perception, and trust. We conduct user studies with six state-of-the-art explainability visualizations, carefully selected to maximize coverage of our taxonomy: SHAP waterfall plots, SHAP force plots, ELI5, LIME, CP profiles, and Anchors (see Section 5 for study details and Section 5.2.1 for visualization descriptions). Prior work has shown that model descriptions and transparency can affect perceived model fairness and trustworthiness [28, 96, 102].

In this section, we analyze how each visualization relates to comprehension, perceived bias, and trust. To provide additional insight, we also consider higher-order connections between these properties. We conclude with a qualitative analysis of free-response answers. Overall, we find that visualization design significantly affects viewer comprehension, bias perception, and trust. Notably, higher comprehension is associated with lower trust. On investigation, we find that this relationship is mediated by bias perception—when the underlying model is biased, people are less likely to trust it.

***Comprehension* and Visualization Design.** To assess how design characteristics influence comprehension, we fit a linear model using the comprehension score defined in Section 5.1. A one-way ANOVA test indicates that visualization type is a significant predictor of comprehension ($p < 0.001$). Participants achieved the highest comprehension scores with LIME, and the lowest with Anchors (emmeans: LIME = 41.2, SHAP waterfall = 39.7, SHAP Force = 38.7, ELI5 = 36.3, CP = 23.9, Anchors = 23.0). Visualizations that explicitly showed the magnitude and direction of feature impacts had higher comprehension (vs. those where they had to be inferred, 38.97 vs. 23.46). This has direct implications for visualization design (see Section 9), and we test this finding causally in RQ3 (see Section 7.1).

Breaking down the comprehension score (Figure 3), we find that visualization is a significant predictor for each component (C1: $p = 0.002$, C2: $p < 0.001$, C3: $p < 0.001$). Participants with LIME visualizations were significantly more likely to correctly assess the loan decision (emmean=7.73), and determine feature impact direction (emmean=27.8) or magnitude (emmean=5.70). This indicates that LIME visualizations characteristics facilitate comprehension. Participants highlighted LIME visual elements as helpful, including its bar chart with classifier output probabilities (taxonomy dimension D10) and feature impact table (D6, D11, D12, D13, and D16). Our follow-up causality experiments (Section 7.2) include these elements.

Finally, we examine if participants are aware of their own comprehension level. A linear model shows that perceived comprehension predicts objective comprehension ($p < 0.001$). However, this effect is small (Pearson's $r = 0.33$). This underscores the importance of visualizations that facilitate actual, rather than perceived, comprehension; model behavior can be incredibly complex, and non-experts may struggle to recognize gaps in their understanding.

***Bias Perception* and Visualization Design.** Using a linear model, we find that visualization type significantly predicts perceived bias

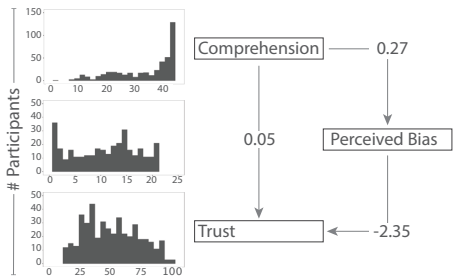


Fig. 5: The mediation effects between comprehension, perceived bias, and trust. While comprehension is negatively correlated with trust, this graph shows that this correlation is heavily mediated by perceived bias. The (estimated) direct effect of comprehension on perceived is significantly positive (0.27), and the direct effect of perceived bias on trust is significantly negative (−2.35), but the estimated direct effect of comprehension on trust is actually slightly positive at 0.05. Furthermore, this last relationship is not statistically significant.

($p = 0.004$, emmeans: LIME = 14.2, SHAP Waterfall = 12.6, SHAP Force = 15.0, ELI5 = 13.9, CP = 10.5, anchors = 11.9). Participants were least likely to perceive bias with CP and Anchors visualizations that have implicit feature impacts, and include alternative outputs (D16, D17, and D22). This indicates that bias perception may be facilitated by simpler visualizations with explicit feature impacts. We use this finding to design a follow-up experiment investigating the impact of manipulating bias perception on trust (see Section 7.3).

***Model Trust* and Visualization Design.** Visualization design also significantly predicts trust ($p = 0.03$). On average, participants trusted CP visualizations most and SHAP force plots least (emmeans: LIME = 50.2, SHAP Waterfall = 52.0, SHAP Force = 45.8, ELI5 = 51.4, CP = 58.3, Anchors= 53.9). The high trust in CP may relate to its implicit design and alternative outputs, while SHAP Force plots' unique horizontal bar elements (D10) may negatively impact trust.

Higher-Order Connections: Comprehension, Bias Perception, and Trust. To provide additional insight, we also consider higher-order connections between comprehension, bias perception, and trust. We observe a small but significant negative correlation between comprehension and trust (Pearson's $r = -0.28$, $p < 0.001$); participants who understood the model better tended to trust it less.

Further analysis reveals this relationship is potentially mediated by bias perception. When plotting the relationship between comprehension, bias, and trust (see Figure 1), we notice a direct inverse relationship between the level of bias perception and trust across tested visualizations. That is, visualizations with higher bias perception also have lower trust. A linear model with trust as the dependent variable and both comprehension and bias perception as the predictors shows that *bias perception alone predicts trust* (Sum sq = 122260, $p < 0.001$), while comprehension is not significant. This is interesting given our finding that comprehension score is negatively correlated with trust.

To better understand this result, we fit a second model with bias perception as the dependent variable, and comprehension as the predictor. In contrast, this model shows comprehension is a significant predictor of bias perception (Sum sq = 3577.3, $p < 0.001$), suggesting that there is potentially a heavy mediation effect of bias perception on the correlation between comprehension score and trust.

To confirm this interpretation, we fit a mediation model with trust score as the dependent variable, the comprehension score as the predictor, and the bias perception score as the mediator. As Figure 5 shows, this analysis reveals that (1) comprehension positively predicts bias perception (estimated coefficient $b = 0.27$, $p < 0.001$), meaning that increasing comprehension by one unit increases perceived bias by 0.27 units on average, (2) bias perception negatively predicts trust ($b = -2.35$, $p < 0.001$), (3) the direct effect of comprehension on trust is small and non-significant ($b = 0.05$, $p = 0.49$), and (4) the indirect effect of comprehension on trust through bias perception is $b = -0.63$.

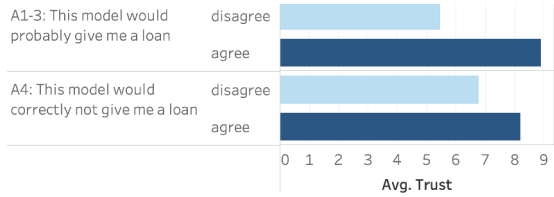


Fig. 6: A bar graph showing the difference in average trust between instances where participants felt the model would give them a loan, and instances where they felt it would not. Average trust here is by instance rather than aggregated. A second set of bars shows the difference in average trust between instances where participants agreed with the statement that the model rightfully would not give them a loan, and instances where they disagreed. Corresponding statements can be found in Figure 3.

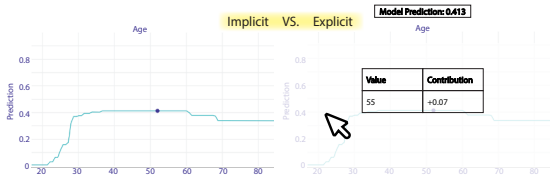
	Comprehension		Fairness	
Dimension	least	most	least	most
Alt. Examples	34	48	33	48
Bars	32	109	33	87
Base Value	52	16	41	7
Color	31	186	26	143
Features	65	65	46	88
Position	9	42	9	30
Values	69	99	61	88
X/Y axis	36	37	35	39

Fig. 7: A bubble chart showing the number of participants mentioning taxonomy features when asked free-form questions regarding the characteristics they found most and least helpful when answering comprehension and fairness questions. These features are grouped by meta-dimension type. Explicitness is excluded as it is not directly mentioned.

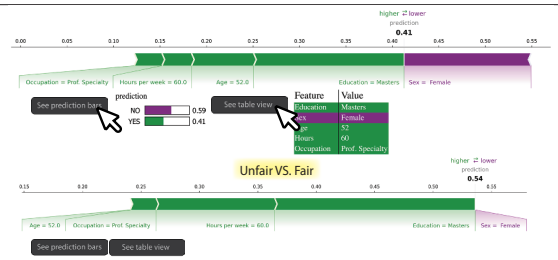
This suggests that the negative link between comprehension and trust is heavily mediated by bias perception: higher comprehension makes bias more apparent, reducing trust. Importantly, this also implies that bias perception—and therefore trust—can potentially be modulated independently of comprehension, motivating the design of visualizations that support both comprehension and bias awareness (see Section 9).

Behavioral Alignment. People are more likely to trust models that benefit them, even if those models are biased [28]. We investigate if participants' relationship to the model (e.g., if they expect it to give them a loan) affected trust. Participants believed the model would give them a loan in about half of all instances (52.15%). In one third (33.86%), participants believed the model would not give them a loan but that this would be the correct choice. Comparing trust scores, we find that participants trusted the model significantly more when they believed it would give them a loan ($p < 0.001$). However, Figure 6 shows that when participants believed they would not receive a loan but agreed with this outcome, there was again a significant (though smaller, 1.54 vs. 3.72) increase in average trust ($p < 0.001$). We define a person holding either of these beliefs about the model to be in behavioral alignment. We find that behavioral alignment has a large positive effect on trust (Cohen's $d = 1.26$) and a large negative effect on bias perception (Cohen's $d = -0.98$). These findings imply that favorable or agreeable outcomes can reduce users' perception of bias and increase trust, regardless of the model's actual fairness.

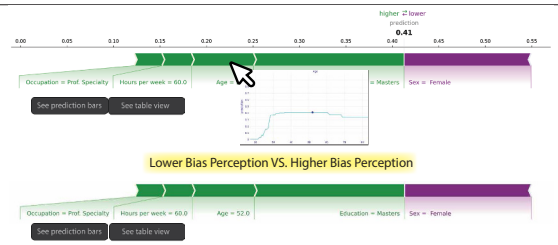
Qualitative Analysis. Finally, we asked participants free response questions regarding which visualization characteristics they found most and least useful, to further inform the design of next-generation explanation visualizations. Via manual analysis, we aggregate responses across taxonomy dimensions. Figure 7 summarizes the number of participants that found each dimension to be most and least helpful when assessing model comprehension or fairness. While participants perceived almost all dimensions as more useful than not, they found color-coding impact direction and magnitude to be particularly useful.



(a) *Experiment 1:* Compares implicit and explicit feature values. We compare responses for a standard CP plot (left) to a CP plot that we manually augmented to show explicit feature values when moused over (right).



(b) *Experiment 2:* Compares responses to an unfair (top) and fair (bottom) model, using a composite visualization designed for high comprehension. The top model uses sex as a significant feature for a loan-recommendation prediction, while the bottom one does not.



(c) *Experiment 3:* Compares responses to composite visualizations facilitating lower (top) and higher (bottom) bias perception. The top visualization adds CP plots on mouse-over of each feature to show alternative outputs and increase visualization complexity, while the bottom removes the x and y axes to increase simplicity.

Fig. 8: Three crowd-sourced controlled experiments testing for causality in between comprehension, bias perceptions, and trust by manipulating comprehension (a), underlying bias (b), and perceived bias (c).

However, anecdotally, for both SHAP Plots, participants found red for positive contributions and blue for negative contributions to be counter-intuitive. Furthermore, both bars and explicit values were found to be useful. Conversely, numbered x/y axes were not found to be particularly useful in conveying information—participants preferred simpler visual cues like color or bar size. Very few participants found base values useful. These qualitative observations motivate the design of our second crowd-sourced controlled experiment in RQ3 (Section 7.2).

RQ2 Summary: Visualization Design Correlations

We find evidence suggesting that comprehension, trust, and bias perception are all affected by visualization design. For example, viewers trusted the underlying model less when given explanations with more explicit information. We further find a negative correlation between comprehension and trust, which is heavily mediated by the perception of bias; when people understand a model more, they trust it less, potentially due to increased bias perception.

7 RQ3: Confirmation of Causality

In RQ2 (Section 6), we observed a correlation between comprehension and trust, mediated by perceived model bias. We also observed that explicit visualization of feature impacts improves comprehension and bias perception. To test the causal nature of these relationships, we conduct three crowd-sourced controlled experiments using the 7 prediction instances from Section 5.2. We use our taxonomy's relevant visualization characteristics to create pairs of identical visualizations, except for the key attribute under test (see Figure 8 for experimental visualization pairs). This approach allows us to establish causality and test if observed correlations generalize beyond specific tools.

Experiment 1—Explicitness: We investigate the explicit feature impact of magnitudes and direction on comprehension and bias perception. This experiment is motivated by our finding in RQ2 that visualizations with explicit information about feature impact and direction had higher comprehension scores. To compare explicit and implicit visualizations, we compare variations of the CP survey described in Section 5.2.1 with and without mouse-over explicit values (See Figure 8a). Given our observations in Section 6, we hypothesize that adding explicitness will increase comprehension and bias perception, while reducing trust.

Experiment 2—Fairness: We test if reducing model bias increases trust by lowering perceived bias, motivated by our findings in RQ2 that bias perception mediates the negative relationship between comprehension and trust. We hypothesize that for high-comprehension visualizations, reducing model bias will also reduce perceived bias and increase trust. We construct an adjustable composite visualization combining design characteristics associated with comprehension (e.g., LIME's probability bar and feature impact table) and bias perception (e.g., SHAP Force's bar chart), which users can explore through interactive mouse-overs.

We then conduct two surveys with this composite visualization, one where participants see only the fair model, and one where participants see only the biased model (see Figure 8b). We chose to reduce bias for sex and age as these were the two protected characteristics from the existing set. We found that our original model gave a loan to 24% of men in a test set, and only 1.4% of women. Furthermore, our model gave 27% of those over the age of 37 (median age) a loan vs. only 5.4% of those under the age of 37. To create our fair model, we used the Seldonian Algorithm [88], constraining model behavior to have demographic parity ($Pr(Y|Group1) - Pr(Y|Group2) < \varepsilon$) for both sex and age. Using the demographic parity metric allowed us to reduce bias in a way that vas visible in individual outputs and did not require a fair training dataset [88]. Using the Seldonian toolkit, we trained a random forest model that fits our fairness constraints while preserving accuracy (accuracy of 0.790 vs. 0.814 for the original model).

Experiment 3—Bias Perception: Finally, we test if altering bias perception through visualization design can affect trust without changing model behavior. We adjust our composite visualization from Experiment 2 to decrease and increase bias perception, creating two new variations (see Figure 8c). To decrease bias perception, we incorporate the characteristics identified in Section 6 as associated with the lowest bias perception by adding CP Plots on feature mouse-over. To increase bias perception, we maximize simplicity by removing characteristics qualitatively identified as unhelpful, such as the x and y axes. We hypothesize that reducing bias perception will increase trust, even if the model's biased behavior does not actually change.

7.1 Experiment 1—Explicitness

We test whether making feature impact magnitudes and directions explicit affects comprehension, bias perception, and trust. Since our distributions for comprehension, perceived bias, and trust are not normal (see Figure 5), we use Wilcoxon Rank sum tests for comparisons and Cohen's d for effect size.

We find that adding explicit indicators of model output, feature impact direction, and feature impact magnitude has a significant effect on all three measures ($p < 0.001$). Specifically, explicitness has a positive medium effect on comprehension ($d = 0.68$), a small positive effect on bias perception ($d = 0.22$), and a small negative effect on trust ($d = -0.26$). This indicates that including explicit values increases comprehension and bias perception, and decreases trust. This result is exciting; not only does it demonstrate how our taxonomy can be used to learn deeper visualization insights, it also indicates that modulating taxonomized design characteristics can increase viewer comprehension of the underlying model, reveal model biases, and adjust viewer trust.

7.2 Experiment 2—Fairness

We test if reducing model bias increases trust, in the context of a high-comprehension visualization. We create this visualization by combining elements of visualizations with high comprehension scores. We use a purple and green color scheme, modifying counterintuitive colors and ensuring they are color-blind friendly. We see significant differences between participants who saw explanations of our unfair vs. fair models, across all measures including comprehension ($p = 0.003$), trust ($p < 0.001$), and bias perception ($p < 0.001$). Using the fair model results in both a small negative effect on bias perception ($d = -0.33$), and also a small-medium positive effect on trust ($d = 0.44$). While significant, the effect on comprehension is negligible ($d = -0.16$).

These findings indicate that with high comprehension, decreasing true bias decreases bias perception, leading to an appropriate increase in trust. This result supports a direct causal relationship between bias perception and trust, and demonstrates that visualizations that facilitate comprehension also calibrate bias perception with true fairness, underscoring the importance of comprehension to visualization design.

7.3 Experiment 3—Bias Perception

Finally, we test if we can use visualization design to modify perceived bias and affect trust, even when the model's underlying behavior is unchanged. We observe significant differences in perceived bias ($p < 0.001$) and trust ($p < 0.001$) between participants who saw our two survey variations. By adjusting our composite visualization to include taxonomy characteristics correlated with decreased bias, we were able cause both a small decrease in bias perception ($d = -0.27$) and also a small increase in trust ($d = 0.20$). There was also a significant change in comprehension ($p < 0.01$), but the effect size was tiny ($d = -0.07$).

These findings indicate that even in the case where changes in comprehension are negligible, visualizations designed to obscure model bias will increase viewer trust in that model. These observations support our hypothesis of a direct causal relationship between bias perception and trust, which can be manipulated through visualization design.

RQ3 Summary: Causal relationships

We find that increasing comprehension of a biased visualization leads to increased bias perception and decreased trust, while high comprehension of a fair visualization results in decreased bias perception and increased trust. However, even when underlying bias does not change, artificially decreasing bias perception can increase trust. These findings support causality for our observed results from RQ2.

8 Limitations

We operationalize our three major measures as described in Section 5.1 based on existing work and the characteristics present in the chosen visualization set. However, there are other methods of measuring comprehension and trust, in particular. Possible comprehension questions will always depend in part on the information intended to be conveyed by the visualization designers. For CP, for instance, feature importance is not an inherent part of designer intent [13]. Therefore, there may be additional questions that better capture comprehension of this visualization that we were unable to ask because the same information was not present in LIME, SHAP, or ELI5. Trust in a model can be operationalized differently as well—for instance, through decision questions [41] or trust games [28]. These measures may more accurately capture trust.

Our model included 6 input features to limit respondent fatigue, which could reduce participant engagement [43]. However, real-world classifier applications likely require larger feature input sets and can potentially involve numerous and overlapping biases. Furthermore, our primary model bias was gender-based, and gender-based AI bias is a well-documented social issue [39], so it is possible that participants were primed to perceive the model as biased upon seeing gender as a feature. Scaling our experiments to more complex models or less widely recognized biases may require alternative survey methodology.

The explanation types of local explainability visualizations vary (see dimensions D24–D27 in Figure 2). For example, SHAP, LIME, and ELI5 present additive feature contribution values, while Anchors presents if-then rules, and CP profiles present outcome sensitivity to individual input feature changes. The type of information presented may impact both comprehension and perception of bias, and offering multiple explanation types in combination may result in a deeper understanding of model functionality. Further research is necessary to better understand how different explanation types both individually and in combination can impact viewers of explainability visualizations.

We neither varied the model's level of bias, beyond creating one fair and one unfair version, nor the model's comprehensibility. Future work investigating more granular bias variations may provide more nuanced insights into the relationship between comprehension and trust.

9 Discussion and Design Implications

We find that the types of visualization characteristics used to impart local explainability information affect a user's comprehension of the underlying model, their perception of bias in that model, and their trust of that model. Explicitly indicating feature importance information, via color or printed values, can increase both comprehension and perception of bias. Qualitative results show that people prefer certain characteristics over others when trying to comprehend model behavior and perceive model bias, and that people are more likely to trust a model they feel would benefit them. Quantitatively, we find a negative correlation between participants' comprehension of and trust in ML models, strongly mediated by the perception of bias. In other words, when dealing with biased models, better visualizations lead people to more greatly perceive bias, reducing trust. However, we also found that certain design decisions can alter bias perception without affecting comprehension. Anecdotally, we notice that visualizations with lower complexity and more explicitness correlate with higher bias perception and lower trust, and this correlation should be explored in future work. Furthermore, the very presence of input features like sex and age resulted in some participants seeing the model as discriminatory, even in the case of the fair model. Our findings indicate that explainability visualizations, when carefully designed, can be a useful tool in revealing ML model behavior and bias to a variety of users, including non-experts. We caution explainability designers to consider the clarity and intuitiveness of their designs to variable user populations, as well as the potential for those designs to obscure problematic model behavior. We further suggest that AI developers looking to use explainability to debug bias in their models consider how different presentation methods may affect their understanding of their own models' functionality.

Anecdotally, all 19 non-binary participants across survey variations found both fair and biased models to be discriminatory. Marginalized communities are more sensitive to discrimination [55], but there is a dearth of research into non-binary individuals' perception of ML bias. Our use of a dataset and model with only two genders may have made non-binary participants feel excluded, possibly affecting their perception of the model. Unfortunately, real-world datasets with data for non-binary individuals are not readily available. Future work should look explicitly at non-binary individuals' perception of bias in technology.

10 Contributions and Future Work

ML-powered systems are increasingly common, but they are often biased. Explanation visualizations may help non-ML-expert stakeholders understand and assess model outputs, but there is a limited understanding of how visualization design can systematically impact user perception. We take steps towards improving our understanding of how explainability visualization design impacts ML model comprehension, bias perception, and trust. First, we survey local explainability visualizations, to create a taxonomy of visualization design characteristics. We then conduct a series of user studies, identifying correlational and causal relationships regarding how these characteristics facilitate comprehension, bias perception, and trust. Our results provide insights for next-generation visualization tools that can better empower stakeholders to make well-informed, responsible decisions about ML applications. Our work forms an important step towards understanding how people's bias perception of ML outputs affects their trust, and how visualization techniques can help improve effectively communicating important aspects of ML models to non-expert, everyday users.

ACKNOWLEDGMENTS

This work is supported by the National Science Foundation under grant no. CCF-2210243.

REFERENCES

[1] D. Ahn et al. Impact of model interpretability and outcome feedback on trust in AI. In *CHI*, 2024. doi: 10.1145/3613904.3642780 3
[2] Y. Ahn and Y.-R. Lin. Fairsight: Visual analytics for fairness in decision making. *TVCG*, p. 1–1, 2019. doi: 10.1109/tvcg.2019.2934262 2
[3] Y. Alufaisan et al. Does explainable artificial intelligence improve human decision-making? arxiv.org/abs/2006.11194, 2020. 3
[4] J. Angwin et al. Machine bias. *ProPublica*, 2016. tinyurl.com/56pfaa6m. 1
[5] V. Arya et al. One explanation does not fit all: A toolkit and taxonomy of AI explainability techniques. arxiv.org/abs/1909.03012, 2019. 3
[6] J. P. Bagrow. Democratizing AI: Non-expert design of prediction tasks. *PeerJ Computer Science*, 6:e296, 2020. doi: 10.7717/peerj-cs.296 2
[7] H. Baniecki et al. dalex: Responsible machine learning with interactive explainability and fairness in Python. *JMLR*, 22(214):1–7, 2021. 2, 5, 6
[8] G. Bansal et al. Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In *CHI*, 2021. doi: 10.1145/3411764.3445717 2, 3
[9] O. Bastani, C. Kim, and H. Bastani. Interpreting blackbox models via model extraction. arxiv.org/abs/1705.08504, 2019. 2
[10] B. Becker and R. Kohavi. Census income. archive.ics.uci.edu/dataset/2/adult, 1996. 5
[11] U. Bhatt, M. Andrus, A. Weller, and A. Xiang. Machine learning explainability for external stakeholders. *arXiv:2007.05408*, 2020. 1, 2
[12] P. Biecek. DALEX: Explainers for complex predictive models in R. *JMLR*, 19(84):1–5, 2018. 2
[13] P. Biecek and T. Burzykowski. *Explanatory Model Analysis*. Chapman and Hall/CRC, New York, 2021. 2, 3, 5, 9
[14] A. Blanco-Justicia and J. Domingo-Ferrer. Machine learning explainability through comprehensible decision trees. In *MLKE*, 2019. 2
[15] J. Buolamwini and T. Gebru. Gender Shades: Intersectional accuracy disparities in commercial gender classification. In *FAccT*, vol. 81, pp. 77–91. PMLR, 2018. 1
[16] N. Burkart and M. F. Huber. A survey on the explainability of supervised machine learning. *JAIR*, 70:245–317, 2021. doi: 10.1613/jair.1.12228 2, 3
[17] Á. A. Cabrera et al. FairVis: Visual analytics for discovering intersectional bias in machine learning. In *VAST*, pp. 46–56, 2019. doi: 10.1109/VAST47406.2019.8986948 1, 2
[18] J. Colin, T. Fel, R. Cadene, and T. Serre. What I cannot predict, I do not understand: A human-centered evaluation framework for explainability methods. arxiv.org/abs/2112.04417, 2023. 3
[19] D. Collaris and J. Wijk. Comparative evaluation of contribution-value plots for machine learning understanding. *Journal of Visualization*, 25:47–57, 2021. doi: 10.1007/s12650-021-00776-w 3, 4
[20] J. W. Crandall et al. Cooperating with machines. *Nature Communications*, 9(1), 2018. doi: 10.1038/s41467-017-02597-8 2
[21] R. J. Crouser et al. Building and eroding: Exogenous and endogenous factors that influence subjective trust in visualization. In *IEEE VIS*, pp. 306–310, Oct. 2024. doi: 10.1109/VIS55277.2024.00069 3
[22] W. K. Diprose et al. Physician understanding, explainability, and trust in a hypothetical machine learning risk calculator. *JAMIA*, 27(4):592–600, 2020. doi: 10.1093/jamia/ocz229 2
[23] U. Ehsan, Q. V. Liao, M. Muller, M. O. Riedl, and J. D. Weisz. Expanding explainability: Towards social transparency in ai systems. In *CHI*, 2021. doi: 10.1145/3411764.3445188 3
[24] H. Elhamdadi, A. Stefkovics, J. Beyer, E. Moerth, H. Pfister, C. X. Bearfield, and C. Nobre. Vistrust: A multidimensional framework and empirical study of trust in data visualizations. *TVCG*, 30(1):348–358, Jan. 2024. doi: 10.1109/TVCG.2023.3326579 3
[25] European Parliament and Council of the European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council. data.europa.eu/eli/reg/2016/679/oj, 2016. 2
[26] Executive Office of the President. Big data: A reporton algorithmic systems, opportunity, and civil rights. tinyurl.com/u4hjp587, 2016. 1
[27] S. A. Friedler, C. Scheidegger, and S. Venkatasubramanian. The (im)possibility of fairness: Different value systems require different mechanisms for fair decision making. *CACM*, 64(4):136–143, Mar. 2021. doi: 10.1145/3433949 2
[28] A. Gaba, Z. Kaufman, J. Cheung, M. Shvakel, K. W. Hall, Y. Brun, and C. X. Bearfield. My model is unfair, do people even care? Visual design affects trust and perceived bias in machine learning. *TVCG*, 30(1):327–337, 2024. doi: 10.1109/TVCG.2023.3327192 1, 2, 3, 4, 6, 7, 9
[29] S. Galhotra, Y. Brun, and A. Meliou. Fairness testing: Testing software for discrimination. In *ESEC/FSE*, pp. 498–510, 2017. doi: 10.1145/3106237.3106277 2
[30] E. D. Gennatas et al. Expert-augmented machine learning. *PNAS*, 117(9):4571–4577, 2020. doi: 10.1073/pnas.1906831117 2
[31] B. Ghai and K. Mueller. D-BIAS: A causality-based human-in-the-loop system for tackling algorithmic bias. *TVCG*, 29(01):473–482, 2023. doi: 10.1109/TVCG.2022.3209484 2, 3
[32] A. Goldstein, A. Kapelner, J. Bleich, and E. Pitkin. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *JCGS*, 24(1):44–65, 2015. doi: 10.1080/10618600.2014.907095 2
[33] R. Guidotti. Counterfactual explanations and how to find them: Literature review and benchmarking. *Data Mining and Knowledge Discovery*, pp. 1–55, 2022. doi: 10.1007/s10618-022-00831-6 2
[34] T. Hailesilassie. Rule extraction algorithm for deep neural networks: A review. arxiv.org/abs/1610.05267, 2016. 2
[35] P. Hall and D. Atherton. Awesome machine learning interpretability. tinyurl.com/2zaj8amsÆŠ, 2024. 3
[36] K. Hartwig and C. Reuter. Nudging users towards better security decisions in password creation using whitebox-based multidimensional visualisations. *Behaviour & Information Technology*, 41(7):1357–1380, 2022. doi: 10.1080/0144929X.2021.1876167 2
[37] G. He, L. Kuiper, and U. Gadiraju. Knowing about knowing: An illusion of human competence can hinder appropriate reliance on ai systems. In *CHI*, 2023. doi: 10.1145/3544548.3581025 3
[38] A. Heimerl, K. Weitz, T. Baur, and E. André. Unraveling ML models of emotion with NOVA: Multi-level explainable AI for non-experts. *TAFFC*, 13(3):1155–1167, 2022. doi: 10.1109/TAFFC.2020.3043603 1, 2
[39] J. Q. Ho, A. Hartanto, A. Koh, and N. M. Majeed. Gender biases within artificial intelligence and ChatGPT: Evidence, sources of biases and solutions. *Computers in Human Behavior: Artificial Humans*, 4:100145, 2025. doi: 10.1016/j.chbah.2025.100145 9
[40] A. Hoag, J. E. Kostas, B. C. da Silva, P. S. Thomas, and Y. Brun. Seldonian toolkit: Building software with safe and fair machine learning. In *ICSE Demo*, 2023. doi: 10.1109/ICSE-Companion58688.2023.00035 1
[41] M. N. Hoque and K. Mueller. Outcome-explorer: A causality guided interactive visual interface for interpretable algorithmic decision making. *TVCG*, 28(12):4728–4740, 2022. doi: 10.1109/tvcg.2021.3102051 3, 9
[42] S. R. Islam, W. Eberle, S. K. Ghafoor, and M. Ahmed. Explainable artificial intelligence approaches: A survey. arxiv.org/abs/2101.09429, 2021. 2
[43] D. Jeong, S. Aggarwal, J. Robinson, N. Kumar, A. Spearot, and D. S. Park. Exhaustive or exhausting? evidence on respondent fatigue in long surveys. *Journal of Development Economics*, 161:102992, 2023. doi: 10.1016/j.jdeveco.2022.102992 9
[44] B. Johnson and Y. Brun. Fairkit-learn: A fairness evaluation and comparison toolkit. In *ICSE Demo*, 2022. doi: 10.1145/3510454.3516830 2
[45] B. Johnson, Y. Brun, and A. Meliou. Causal testing: Understanding defects' root causes. In *ICSE*, 2020. doi: 10.1145/3377811.3380377 2
[46] B. Johnson et al. Fairkit, fairkit, on the wall, who's the fairest of them all? Supporting data scientists in training fair models. *EURO Journal on Decision Processes*, 11, 2023. doi: 10.1016/j.ejdp.2023.100031 1, 2
[47] Z. Kaufman, M. Endres, C. Xiong Bearfield, and Y. Brun. Supplementary materials. https://osf.io/c87xm/?view_only=31dfc1f2a7624f5cb20b0f07d3730df3. 2, 3, 5
[48] H. Kaur et al. Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. In *CHI*, p. 1–14, 2020. doi: 10.1145/3313831.3376219 3
[49] G. Ke et al. LightGBM: A highly efficient gradient boosting decision tree. In *NeurIPS*, 2017. 5
[50] S. S. Y. Kim, N. Meister, V. V. Ramaswamy, R. Fong, and O. Russakovsky. HIVE: Evaluating the human interpretability of visual explanations. In *ECCV*, p. 280–298, 2022. doi: 10.1007/978-3-031-19775-8_17 3
[51] J. Klaise, A. V. Looveren, G. Vacanti, and A. Coca. Alibi explain: Algorithms for explaining machine learning models. *JMLR*, 22(1), 2021. 2
[52] M. Komorowski, L. A. Celi, O. Badawi, A. C. Gordon, and A. A. Faisal. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature Medicine*, 24(11):1716–1720, 2018. 1, 2

[53] M. Korobov and K. Lopuhin. ELI5. github.com/eli5-org/eli5/tree/master, 2016–2017. 2, 3, 5, 6

[54] S. Krishna et al. The disagreement problem in explainable machine learning: A practitioner's perspective. arxiv.org/abs/2202.01602, 2022. 5

[55] M. K. Lee and K. Rich. Who is included in human perceptions of AI?: Trust and perceived fairness around healthcare AI and cultural mistrust. In *CHI*, 2021. doi: 10.1145/3411764.3445570 3, 9

[56] A. Liang, J. Lu, and X. Mu. Algorithm design: A fairness-accuracy frontier. arxiv.org/abs/2112.09975, 2023. 2

[57] Z. C. Lipton. The mythos of model interpretability. arxiv.org/abs/1606.03490, 2017. 2

[58] M. Liu, J. Shi, K. Cao, J. Zhu, and S. Liu. Analyzing the training processes of deep generative models. *TVCG*, 24(1):77–87, 2018. doi: 10.1109/TVCG.2017.2744938 2

[59] S. M. Lundberg et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature Biomedical Engineering*, 2(10):749, 2018. 2, 3, 5, 6

[60] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In *NeurIPS*, vol. 30, 2017. 2, 3, 5, 6

[61] S. Ma et al. Who should I trust: AI or myself? leveraging human and AI correctness likelihood to promote appropriate trust in AI-assisted decision-making. arxiv.org/abs/2301.05809, 2023. 3

[62] S. Maksymiuk, A. Gosiewska, and P. Biecek. Landscape of R packages for eXplainable AI. arxiv.org/abs/2009.13248, 2021. 3

[63] S. Mertes, T. Huber, K. Weitz, A. Heimerl, and E. André. Ganterfactual—counterfactual explanations for medical non-experts using generative adversarial learning. *Frontiers in AI*, 5:825565, 2022. 1, 2

[64] Microsoft. Lightgbm. github.com/microsoft/LightGBM, 2017. 5

[65] Y. Ming, H. Qu, and E. Bertini. Rulematrix: Visualizing and understanding classifiers with rules. *TVCG*, 25(1):342–352, 2019. doi: 10.1109/TVCG.2018.2864812 2, 3

[66] Y. Nakao et al. Toward involving end-users in interactive human-in-the-loop AI fairness. *ACM TiiS*, 12(3), July 2022. doi: 10.1145/3514258 4

[67] Y. Nakao and Y. Sugano. Use of machine learning by non-expert DHH people: Technological understanding and sound perception. In *NordiCHI*, 2020. doi: 10.1145/3419249.3420157 2

[68] M. Nauta et al. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI. *ACM Computing Surveys*, 55(13s):1–42, 2023. doi: 10.1145/3583558 3

[69] R. C. Nickerson, U. Varshney, and J. Muntermann. A method for taxonomy development and its application in information systems. *European Journal of Information Systems*, 22:336–359, 2013. 3

[70] E. Özalp, K. Hartwig, and C. Reuter. Trends in explainable artificial intelligence for non-experts. *KI-Kritik/AI Critique Volume 4*, p. 223, 2023. 1, 2

[71] S. Palan and C. Schitter. Prolific.ac — A subject pool for online experiments. *JBEF*, 17:22–27, 2018. 4, 6

[72] Posit. RStudio desktop. posit.co/download/rstudio-desktop/, 2025. 5

[73] F. Poursabzi-Sangdeh et al. Manipulating and measuring model interpretability. In *CHI*, 2021. doi: 10.1145/3411764.3445315 3

[74] Q.ai. How intelligent machines are reshaping investing. *Forbes*, 2022. 1, 2

[75] I. Qualtrics. Qualtrics. *Provo, UT, USA*, 2013. 4

[76] A. Rechkemmer and M. Yin. When confidence meets accuracy: Exploring the effects of multiple performance indicators on trust in machine learning models. In *CHI*, 2022. doi: 10.1145/3491102.3501967 3

[77] M. T. Ribeiro, S. Singh, and C. Guestrin. "Why should I trust you?": Explaining the predictions of any classifier. In *KDD*, pp. 1135–1144, 2016. doi: 10.1145/2939672.2939778 2, 3, 5, 6

[78] M. T. Ribeiro, S. Singh, and C. Guestrin. Anchors: High-precision model-agnostic explanations. In *AAAI*, 2018. 2, 3, 5, 6

[79] Y. Rong et al. Towards human-centered explainable AI: A survey of user studies for model explanations. *TPAMI*, 46(4):2104–2122, 2024. doi: 10.1109/TPAMI.2023.3331846 3

[80] P. K. Roy, S. S. Chowdhary, and R. Bhatia. A machine learning approach for automation of resume recommendation system. *Procedia Computer Science*, 167:2318–2327, 2020. 1, 2

[81] I. Salehin et al. AutoML: A systematic review on automated machine learning with neural architecture search. *Journal of Information and Intelligence*, 2(1):52–81, 2024. doi: doi.org/10.1016/j.jiixd.2023.10.002 2

[82] W. Samek, T. Wiegand, and K.-R. Müller. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. arxiv.org/abs/1708.08296, 2017. 2

[83] N. Sheard and A. Schwartz. The movement to ban government use of face recognition. tinyurl.com/56tfjkjs, 2022. 1

[84] H. Shen and T.-H. K. Huang. How useful are the machine-generated interpretations to general users? A human evaluation on guessing the incorrectly predicted labels. In *HCOMP*, 2020. 3

[85] N. Singer. Amazon faces investor pressure over facial recognition. *NYT*, 2019. 1

[86] M. Staniak and P. Biecek. Explanations of model predictions with live and breakDown packages. *The R Journal*, 2018. doi: 10.32614/RJ-2018-072 2

[87] I. Stepin, J. M. Alonso, A. Catala, and M. Pereira-Fariña. A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. *IEEE Access*, 9:11974–12001, 2021. doi: 10.1109/ACCESS.2021.3051315 2

[88] P. S. Thomas, B. C. da Silva, A. G. Barto, S. Giguere, Y. Brun, and E. Brunskill. Preventing undesirable behavior of intelligent machines. *Science*, 366(6468):999–1004, 2019. doi: 10.1126/science.aag3311 1, 2, 8

[89] S. Tolmeijer, M. Christen, S. Kandul, M. Kneer, and A. Bernstein. Capable but amoral? Comparing AI and human expert collaboration in ethical decision making. In *CHI*, 2022. doi: 10.1145/3491102.3517732 3

[90] N. van Berkel, J. Goncalves, D. Russo, S. Hosio, and M. B. Skov. Effect of information presentation on fairness perceptions of machine learning predictors. In *CHI*, 2021. doi: 10.1145/3411764.3445365 2, 3

[91] R. Vanderford. New York's landmark AI bias law prompts uncertainty. *WSJ*, Sept. 21, 2022. tinyurl.com/wv4ueczf. 1

[92] M. Veale and R. Binns. Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society*, 4(2), 2017. 2

[93] D. Wang, W. Zhang, and B. Y. Lim. Show or suppress? Managing input uncertainty in machine learning model explanations. arxiv.org/abs/2101.09498, 2021. 2

[94] Q. Wang, K. Huang, P. Chandak, M. Zitnik, and N. Gehlenborg. Extending the nested model for user-centric xai: A design study on gnn-based drug repurposing. *TVCG*, PP, 10 2022. doi: 10.1109/TVCG.2022.3209435 3

[95] Q. Wang, Z. Xu, Z. Chen, Y. Wang, S. Liu, and H. Qu. Visual analysis of discrimination in machine learning. *TVCG*, 27(2):1470–1480, 2021. doi: 10.1109/TVCG.2020.3030471 2, 4

[96] R. Wang et al. Factors influencing perceived fairness in algorithmic decision-making: Algorithm outcomes, development procedures, and individual differences. In *CHI*, 2020. doi: 10.1145/3313831.3376813 2, 3, 6

[97] J. Wexler et al. The what-if tool: Interactive probing of machine learning models. *TVCG*, p. 1–1, 2019. doi: 10.1109/tvcg.2019.2934619 2

[98] P. Wielopolski, O. Furman, J. Stefanowski, and M. Zieba. Unifying perspectives: Plausible counterfactual explanations on global, group-wise, and local levels. https://arxiv.org/abs/2405.17642, 2024. 2

[99] T. Xie, Y. Ma, J. Kang, H. Tong, and R. Maciejewski. FairRankVis: A visual analytics framework for exploring algorithmic fairness in graph mining models. *TVCG*, 28(1), 2022. doi: 10.1109/TVCG.2021.3114850 2

[100] C. Xiong et al. Illusion of causality in visualized data. *TVCG*, 26(1):853–862, 2020. doi: 10.1109/TVCG.2019.2934399 2, 3

[101] C. Xiong et al. Reasoning affordances with tables and bar charts. *TVCG*, pp. 1–13, 2022. doi: 10.1109/TVCG.2022.3232959 2, 3

[102] C. Xiong, L. Padilla, K. Grayson, and S. Franconeri. Examining the components of trust in map-based visualizations. In *TrustVis*, 2019. 6

[103] A. Yala, C. Lehman, T. Schuster, and T. P. andRegina Barzilay. A deep learning mammography-based model for improved breast cancer risk prediction. *Radiology*, 2019. doi: 10.1148/radiol.2019182716 1

[104] F. Yang et al. Swaying the public? Impacts of election forecast visualizations on emotion, trust, and intention in the 2022 U.S. midterms. *TVCG*, 30(01):23–33, 2024. doi: 10.1109/TVCG.2023.3327356 3

[105] F. Yang, Z. Huang, J. Scholtz, and D. L. Arendt. How do visual explanations foster end users' appropriate trust in machine learning? In *IUI*, pp. 189–201, 2020. doi: 10.1145/3377325.3377480 3

[106] Q. Yang, J. Suh, N.-C. Chen, and G. Ramos. Grounding interactive machine learning tool design in how non-experts actually build models. In *DIS*, pp. 573–584, 2018. doi: 10.1145/3196709.3196729 2

[107] M. Yin, J. Wortman Vaughan, and H. Wallach. Understanding the effect of accuracy on trust in machine learning models. In *CHI*, pp. 1–12, 2019. doi: 10.1145/3290605.3300509 3

[108] Y. Zhang, Q. V. Liao, and R. K. E. Bellamy. Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In *FAT**, p. 295–305, 2020. doi: 10.1145/3351095.3372852 3