# Do Explanations Improve the Quality of AI-assisted Human Decisions? An Algorithm-in-the-Loop Analysis of Factual & Counterfactual Explanations

Lujain Ibrahim [1]
New York University Abu Dhabi
Abu Dhabi, United Arab Emirates
lujain.ibrahim@nyu.edu

Mohammad M. Ghassemi
Michigan State University
East Lansing, United States
ghassem3@msu.edu

Tuka Alhanai
New York University Abu Dhabi
Abu Dhabi, United Arab Emirates
tuka.alhanai@nyu.edu

## ABSTRACT

The increased use of AI algorithmic aids in high-stakes decision making has prompted interest in explainable AI (xAI), and the role of counterfactual explanations to increase trust in human-algorithm collaborations and to mitigate unfair outcomes. However, research is limited in understanding how explainable AI improves human decision-making. We conduct an online experiment with 559 participants, utilizing an “algorithm-in-the-loop" framework and real-world pre-trial data to investigate how explanations of algorithmic pretrial risk assessments generated from state-of-the-art machine learning explanation methods (counterfactual explanations via DiCE & factual explanations via SHAP) influences the quality of decision-makers' assessment of recidivism. Our results show that counterfactual and factual explanations achieve different desirable goals (*separately* improve human assessment of model accuracy, fairness, and calibration), yet still fall short of improving the *combined* accuracy, fairness, and reliability of human predictions — reinstating the need for sociotechnical, empirical evaluations in xAI. We conclude with user feedback on DiCE counterfactual explanations, as well as a discussion of the broader implications of our results to AI-assisted decision-making and xAI.

## KEYWORDS

explanations; counterfactuals; risk assessments; trust; fairness; sociotechnical systems; user studies

**ACM Reference Format:**
Lujain Ibrahim [1], Mohammad M. Ghassemi, and Tuka Alhanai. 2023. Do Explanations Improve the Quality of AI-assisted Human Decisions? An Algorithm-in-the-Loop Analysis of Factual & Counterfactual Explanations. In *Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), London, United Kingdom, May 29 – June 2, 2023*, IFAAMAS, 9 pages.

## 1 INTRODUCTION

Explainable Artificial Intelligence (xAI) has become an essential part of the “responsible AI" agenda set by both academia and industry, garnering great interest among stakeholders from decision-makers and decision-subjects, to regulatory bodies and engineers [6, 19]. Indeed, critical life-changing decisions are being undertaken by decision-makers with the assistance of AI; one-third of counties in the United States (U.S.) utilize algorithmic risk assessment tools to inform pretrial release/detention of defendants pending adjudication of their cases [9]. In light of this, model *explanations* have been identified as a promising locus for promoting accountability, detecting bias, and promoting human-algorithm trust [10, 28, 38].

This interest in xAI has led to substantial technical advances in the explainablility of machine learning models, from the development of model-agnostic methods (LIME and Shapley values) to example-based methods (Counterfactual and Adversarial Explanations) [5]. This technical progress, however, has not come without its criticisms; a growing volume of work points to the gap between research on real-world, user-centric needs for AI explanations and the current algorithm-centric work in xAI which relies on “researchers' intuition of what constitutes a ‘good' explanation” [31, 32]. Counter-intuitive findings from user studies on the human-interpretability of current xAI approaches to model explanations (e.g. findings that varying algorithmic explanations does not improve human performance, and that users do not prefer simple explainable models over black box models [35, 39]) also point to the need for more comprehensive empirical evaluations of how explanations achieve their intended performance and accountability goals (if at all).

In this study, we address the gap between algorithm-centric and user-centric work by investigating, in an online experiment with real-world pretrial data, the influence of state-of-the-art machine learning model explanations on human assessments of recidivism within *Green and Chen*'s “algorithm-in-the-loop" framework [18]. It is important to note that we use risk of recidivism prediction merely as an example to test our hypotheses due to its prevalence as an application of AI-assisted decision-making as well as its dataset availability. There is a rich literature, which is outside of the scope of this paper, that outlines the serious limitations and negative impact of using these tools in practice [12, 15, 25, 44].

## 2 CONTRIBUTIONS

The contributions of this study are three fold: (1) we introduce an additional normative principle, “effective explanations," to the “algorithm-in-the-loop" framework; a principle from human-computer interaction research and the explanation sciences, (2) we offer, to the best of our knowledge, the first study with empirical evidence (including user feedback) on different human-algorithm aspects of explanations generated using the Diverse Counterfactual Explanations (DiCE) library [34], and (3) we examine from a variety of angles whether explanations of different types (counterfactual vs factual) and of different characteristics (diverse/complete vs selective), achieve their intended goals of improving accuracy, reliability,

*Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), A. Ricci, W. Yeoh, N. Agmon, B. An (eds.), May 29 – June 2, 2023, London, United Kingdom.*

and fairness of AI-assisted human decisions. We also make our code publicly available. [1]

# 3 RELATED WORK

## 3.1 Types of Explanations in xAI

Recent work in xAI has primarily targeted two aims: (1) transparency and (2) post-hoc explanations. In our study, we focus on post-hoc explanations, particularly local explanations, as they are largely what decision-makers and decision-subjects interact with, especially given the increasing interest in the deployment of deep learning models [3]. Addressing current xAI directions, *Wachter et al.* has posited from the explanation sciences that the xAI community views generating explanations too narrowly, and proposed that human-desirable explanations should be "contrastive, selective, and social" [32]. We test two of these tenets, "contrastive" and "selective," in our study. We explore "contrastive" through presenting counterfactual explanations (via counterfactual examples) versus factual explanations (via feature attribution), and "selective" through selecting what to show in these explanations according to either epistemic (in the counterfactual case) or statistical (in the factual case) importance.

## 3.2 Limited Empirical Studies on Explanations

Previous studies providing quantitative evaluations of the human-interpretability of AI explanations have covered multiple fronts. Some have examined what *types* of explanations are preferable; *Lakkaraju et al.* found through a user study that subjects are faster and more accurate at describing local decision boundaries based on decision sets as opposed to rule lists [24]. Closer to our own work, *Riverio and Thrill* showed that subjects found counterfactual explanations appropriate when their expectations matched the model output, and found neither factual nor counterfactual explanations appropriate when their expectations did not match the model output [41]. Others, like *Kulesza et al.* and *Narayanan et al.* have investigated *properties* of explanations like complexity (number of lines, new concepts), soundness ("nothing but the truth"), and completeness ("the whole truth") and how they affect users' mental models [23, 35]. Similar to these studies, our work provides empirical grounding on the effects of the types and properties of explanations on AI-assisted human decisions.

## 3.3 The "Algorithm-in-the-loop" Framework

We utilize *Green and Chen*'s framework (consisting of normative principles and evaluation metrics) to evaluate what they define as "algorithm-in-the-loop" systems: systems that "employ algorithmic decision making aids to enhance human decision making" [18]. The framework's normative principles are *accuracy*, *reliability*, and *fairness* of human decisions. Unlike the human-in-the-loop paradigm, the framework focuses on a sociotechnical evaluation that centers human (instead of algorithm) performance as the main outcome of interest.

The participants of their two experiments (on AI-assisted risk of recidivism and loan spending predictions) largely failed to satisfy the principles of accuracy, reliability, and fairness. *Green and Chen* also counter-intuitively found that providing participants with explanations or feedback did not improve their performance. Their results exhibited the limitations of such systems and their promise to enhance human decision making into a more efficient and ethical affair. Establishing new principles to evaluate these systems is needed to investigate their limitations and improve or reassess whether their utilization for certain tasks is justified.

# 4 WHAT MAKES AN "EFFECTIVE EXPLANATION"?

In addition to *Green and Chen*'s three principles of accuracy, reliability, and fairness, we identify an additional principle to govern decision-makers' interactions with AI explanations: "effective explanations". We identified the following desiderata for this principle: *trust*, *insight*, and *fair and ethical decision making* [27, 28, 45]:

**Trust**: Explanations should elicit *appropriate* trust in the algorithm, i.e. trust should be calibrated to match the algorithm's performance (e.g. fairness and accuracy) [8, 42].

**Insight**: Explanations should convey useful information about the local prediction and potentially the algorithm's inner-workings (directly via global explanations, or indirectly via observed patterns over many local explanations).

**Fair and Ethical Decision making**: Explanations should mitigate decision-makers' biases, and make decision-makers more apt at identifying and correcting points of limitations in the system (e.g. false positives and false negatives) [45].

# 5 METHODS

This study was approved by the Human Subjects Research Program Institute Review Board (IRB) at New York University Abu Dhabi.

## 5.1 Data, Model, & Explanations

### 5.1.1 Dataset.

We focused our experiment on pretrial detention as machine learning is increasingly being used in courts around the U.S. to predict pretrial violation [9]. We utilized ProPublica's public, real-world dataset from their 2016 investigation into the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) risk assessment tool used by many U.S. courts [1]. The dataset included COMPAS risk scores for likelihood of recidivism, on a scale from 0 to 10 (at increments of 1), for pretrial defendants. Our processed dataset consisted of 6,159 defendants, 34% of which were Caucasian, 51.4% of which were African-American, and the rest of which were Hispanic, Asian, Native American, or other.

### 5.1.2 Risk Assessment Model.

We trained a gradient boosted classifier to predict risk of recidivism using defendants' age, number of prior convictions, number of juvenile felony charges, number of juvenile misdemeanor charges, current offense type, and current charge degree (felony/misdemeanor) [14]. Since race and gender are normally excluded in the training of real-world risk assessment tools, they were also excluded from our model [2]. With an 80-20 train-test split, our model achieved an area under the receiver operating curve of 0.68 — a result comparable to COMPAS and the Public Safety Assessment [30, 36]. When assessing our model's fairness, we found our model to be well-calibrated. 300 defendants from the test set were then randomly selected for use in the study.

[1] https://github.com/x-labs-xyz/aamas23-factual-counterfactual-explanations

*5.1.3 SHAP & DiCE Explanations.* To obtain factual explanations in the form of feature attribution explanations, we employed SHapley Additive exPlanations (SHAP) to generate shapley values that show what features (and by how much) push the model output to be higher or lower than the average model output over the training data [29]. The features were then divided into two sets: those that push the model output to be higher (positive shapley values) than the average model output, and those that push it to be lower (negative shapley values). Each set was then sorted in descending order (i.e. in order of importance).

To generate counterfactual explanations, we employed the DiCE library to obtain three counterfactual examples for each defendant [34]. These examples were randomly generated and "diverse" in the sense that they offered a number of different changes across different features needed to generate the opposite model output. DiCE allows setting constraints on the values of features as well as the type of features that can be varied. We constrained the age value to be between 18 and 96, the highest and lowest ages in our dataset. For the treatment with "diverse" counterfactual explanations, we allowed any of the six features used in the model to be varied in the generated examples, but for the treatment with "selective" counterfactual explanations, we only allowed the age, number of prior convictions, and charge degree (felony/misdeameanor) to be varied. These three features were chosen as they were deemed to carry more weight in the defined risk formula released by the Public Safety Assessment, which we considered an indication of the epistemic importance of those features in the prediction of recidivism [2]. When the risk assessment model score was greater than 5 (high risk of recidivism), counterfactual examples were presented that would reverse that outcome, i.e. examples that show changes to the defendant profile that would make the algorithm consider the defendant low risk. Similarly, when the score was less than or equal to 5 (low/medium risk), examples were presented that would make the algorithm consider the defendant high risk.

## 5.2 Experiment Task

Each experiment session consisted of a tutorial page, a demographic survey, 30 prediction tasks of the same treatment, and an exit survey, in that order.

*5.2.1 Tutorial.* The tutorial, accessible to participants throughout their session, consisted of explanations of criminal justice terminology (what pretrial detention is, examples of crimes under each category, and what different parts of the defendant profile mean), and risk assessments (what the risk assessment model is predicting, which parts of the defendant profile it uses, and that "As with most algorithms, the predictions are not 100% accurate.").

*5.2.2 Intro Survey.* The experiment opened with eight multiple choice questions on participants' gender, age, education degree level, ethnicity, political views, technical literacy (machine learning familiarity), and domain literacy (U.S. Criminal Justice familiarly) [17]. The literacy questions were on a 5-point Likert scale.

*5.2.3 Task Layout.* Each prediction task presented participants with the profile of a defendant, along with, depending on the treatment participants were assigned to, either the risk assessment model score only or the model score and an explanation (factual or counterfactual) to shed light on why the model predicted that score. The defendant profile included the six features that the model was trained on along with the race and gender of defendants to mimic the information judges are presented with [11]. A profile example can be seen in Appendix Figure 1. Participants were then asked to provide a risk score, by choosing from options on a scale from 0 to 10 at increments of 1, for each defendant: "How likely is this person to commit another crime before trial?"

*5.2.4 Treatments.* Participants were presented 30 prediction tasks from one of the following six treatments:

(1) **Baseline:** only the defendant profile with no reference to the risk assessment model. This was a control treatment.
(2) **Risk Assessment Model Only (unexplained):** the defendant profile along with the decile risk assessment model score with no further explanation. This was also a control treatment.
(3) **Diverse Counterfactual:** the defendant profile, the decile risk assessment model score, and three "diverse" counterfactual explanations presenting variations to the defendant profile that would flip the risk assessment model's predicted outcome.
(4) **Selective Counterfactual** the defendant profile, the decile risk assessment model score, and three "selective" counterfactual explanations presenting variations to the profile that would flip the risk assessment model's predicted outcome. "Selective" was defined in terms of selecting what features could be varied in the explanations.
(5) **Complete Feature Attribution:** the defendant profile, the decile risk assessment model score, and a list of features that make the defendant higher risk and a list of features that make them low risk compared to the average model output, in descending order of importance.
(6) **Selective Feature Attribution:** the defendant profile, the decile risk assessment model score, and only the most influential feature that makes the defendant higher risk, and only the most influential feature that makes them lower risk compared to the average model output.

*5.2.5 Exit Survey.* The experiment concluded with eight multiple choice questions (on a 5-point Likert scale) and two open-response questions on participants' response confidence levels, perceptions of algorithm accuracy and fairness, use of presented explanation, ability to explain their decision making process, and level of accountability they should face (relative to the algorithm's developers) if their decision is contested [16].

## 5.3 AMT Data Collection & Processing

Experiments were conducted using AMT human intelligence tasks (HITs). AMT workers were U.S.-based, previously completed at least 1,000 HITs with an 85%+ approval rating, [21]. The intro and exit surveys included one attention check each [4].

## 5.4 Evaluation Metrics & Models

We used the following metrics defined by *Green and Chen*[2]:

*5.4.1 Prediction Score & Gain.* The brier score was used to assess the quality of participant predictions, defined as: $(f-o)^2$ where $f$ is the predicted (forecasted) probability, and $o$ is the outcome (1 if the event occured, and 0 if it didn't occur). A brier score may take any value between 0 and 1; the lower the brier score, the more accurate the predictions, hence a brier score of 0 is the best, and 1 is the worst. From the brier score, a *prediction score* for each participant was calculated using the average brier score of the 30 predictions a participant made in their treatment, and is defined as: $\frac{1}{N}\sum_{i=0}^{N}(1-(f_i-o_i)^2)$ where $N$ is the number of predictions (i.e. 30) for a set of $i$ tasks. The prediction scores for the risk assessment model was calculated in the same way. From the prediction score, the *average* prediction score $S$ across all participants in a given treatment $t$ may be calculated as: $S_t = \frac{1}{M}\sum_{j=0}^{M}\frac{1}{N}\sum_{i=0}^{N}(1-(f_i-o_i)^2)$ where $M$ is the total number of participants in the treatment.

Additionally, the performance *gain* achieved by each treatment ($Gain_t$) was defined as the ratio between (a) $\Delta_t$: the average *prediction score* of participants in the treatment over participants in the baseline ($S_t - S_B$), and (b) $\Delta_R$: the performance of the risk assessment model over the participants in the baseline ($S_R - S_B$):

$$Gain_t = \frac{\Delta_t}{\Delta_R} = \frac{S_t - S_B}{S_R - S_B}$$

where $S_t$, $S_B$, $S_R$ are the average *prediction scores* of participants in treatment $t$, participants in the baseline $B$, and the risk assessment model $R$, respectively.

*5.4.2 Influence.* To assess how much participants altered their predictions when presented with different types of information (i.e. treatments), we calculated an *influence* score. The influence of the risk assessment model on prediction $p_i^k$ by participant $k$ about defendant $i$ was defined as:

$$I_i^k = \frac{p_i^k - b_i}{r_i - b_i}$$

where $b_i$ is the average prediction made about the defendant by participants in the baseline treatment and $r_i$ is the prediction made about the defendant by the model. To obtain reliable calculations, predictions for which $|r_i - b_i| < 0.05$ were excluded [17]. This metric allowed us to evaluate differences between individual predictions, and not just averages (i.e. average prediction scores). $I = 0$ means the participant ignored the risk assessment model, $I = 0.5$ means the participant equally weighted their initial prediction and risk assessment model, and $I = 1$ means the participant completed relied on the model for their prediction. For each defendant in the dataset, the average influence the model had on participants making predictions about that defendant was also calculated across all treatments as well as in each individual treatment.

*5.4.3 Disparate Interactions.* Participants' "disparate interactions" or racially-biased interactions with the risk assessment model were evaluated in two ways. Firstly, using the influence metric, by calculating the *influence disparity* between black and white defendants in two cases: (1) when the model prediction was higher than the average prediction participants made in the baseline treatment (RA influence disparity$_>$: cases when the model encouraged participants to increase their score), and (2) when the model prediction was lower than the average baseline prediction (RA influence disparity$_<$: cases when the model encouraged participants to decrease their score):

$$\text{RA influence disparity}_{>/<} = I_{black,>/<} - I_{white,>/<}$$

where $I_{black}$ and $I_{white}$ are the average influence values for black and white defendants respectively.

The second measure was *deviation disparity*, where *deviation* was defined as the degree to which participants increased or decreased the model's score for each defendant. Deviation was calculated for each prediction task, and the average deviation for white and black defendants and the deviation disparity between the two races were also calculated:

$$\text{Deviation disparity} = D_{black} - D_{white}$$

where $D_{black}$ and $D_{white}$ are the average deviation for black and white defendants respectively.

*5.4.4 Overview of Statistical Models.* We ran linear, ordinary least squares (OLS) regressions to examine associations between different features, and linear mixed-effects models when analyzing results at the level of predictions to account for repeated samples of defendants and participants in the predictions [22]. T-Tests were used to examine statistical significance.

# 6 RESULTS

## 6.1 Demographics of AMT Participants

There were 738 unique participants who participated in our study for a fee of $2.40. After excluding responses from those who failed either one of the attention checks or had incomplete/poor quality responses, the number of participants was reduced from 738 to 559 participants, resulting in 16,770 predictions available for use in the analyses (559 participants x 30 predictions-per-participant).

Slightly more men (52.4%) than women (47.6%) participated. The mean age was 35-44 years with 27.3% of participants within that age range. More than half of participants reported an education level of a bachelor's degree or higher (72.6%), and a majority self-identified as White (71.3%). On average, in all treatments except the baseline treatment, participants reported a greater familiarity with the U.S. criminal justice system than with machine learning.

## 6.2 Approaches to Prediction Task

Mentions of features that participants took into account when predicting risk scores was common across all treatments where the most referenced features were age, criminal history, and nature of the crime. However, some participants' approaches to race were notable; a few participants expressed concerns with algorithms (in general, not ours specifically), their biased predictions, and their use in criminal justice, one even explicitly mentioning that these concerns guided how they approached our experiment from the get-go, "I took it for granted that the algorithm, programmed

[2] The analyses were run on a standard 2019 MacBook Pro (2.4GHz 8-Core 9th Gen Intel Core i9 Processor, 32GB 2400MHz memory)

by humans, would incorporate a degree of cultural and systemic racism in its judgments. So, given other measures of likelihood of commission of crimes before trial, I tended to downgrade its risk scores of African-American subjects". As these general concerns were in response to an optional, open-response question that less than half of participants completed, the responses indicate that there may have been other participants who also approached the prediction tasks with an awareness of algorithmic bias.

## 6.3 Strategies for Incorporating Explanations

Participants' responses to how they incorporated the explanations could be categorized into the following attitudes: (1) explanations were useful/used in their decision making process, (2) explanations were insightful to finding patterns in how the algorithm works, (3) explanations were not useful due to specific critiques, and (4) explanations were ignored for no specified reason. Examples for each category are available in the Appendix.

**Table 1: *Performance gains of each treatment.* Gains of the factual treatments were larger than those of the counterfactual treatments, and gains of the diverse/complete treatments were larger than those of the selective treatments.**

| Treatment | Prediction Score | Gain |
|---|---|---|
| RA Only | 0.76 | 0.56 |
| Diverse Counterfactual | 0.75 | 0.45 |
| Selective Counterfactual | 0.74 | 0.38 |
| Complete Feature Attr. | 0.76 | 0.54 |
| Selective Feature Attr. | 0.75 | 0.47 |

## 6.4 Principle 1: Accuracy

Across all treatments, the average risk assessment model *prediction score* was 0.80 and the average participant *prediction score* was 0.74. Participants in all five treatments still under-performed the model performance ($p < 10^{-6}$). Even though all four explanation treatments improved ($p < 0.005$) participant performance relative to the baseline (no risk assessment model) treatment, they did not lead to a statistically significant increase in participant performance relative to the unexplained risk assessment model treatment. As seen in Table 1, the performance gains of the explanation treatments were also lower than of the unexplained risk assessment model treatment, where the largest gain in an explanation treatment was a gain of 0.54 in the complete feature attribution treatment.

## 6.5 Principle 2: Reliability

*Green and Chen* defined a reliable prediction as one for which participants (1) accurately evaluated both their own and the algorithm's performance, and (2) calibrated the incorporation of the algorithm's score into their prediction based on its performance. We also introduce (3) an exploration of accountability under this principle. The results are summarized in Table 2.

*6.5.1 Evaluation.* We used two survey responses to look at participants' evaluation of their own performance: (1) self-reported performance confidence, and (2) self-reported performance confidence relative to other participants. The latter was a question included to allow participants to not only report their absolute performance confidence (an incomplete measure due to the largely non-expert participant population), but also their performance confidence relative to other non-experts (other participants completing the experiment). The mean survey responses for both questions were between “Moderately" (3) and “Very" (4) confident for all treatments.

**Table 2: *Reliability of participant predictions.* The reliability results of each treatment (OLS regression between self-report and actual outcome) showing participants' ability to evaluate their own performance (`Conf`), ability to evaluate the risk assessment model accuracy (`Acc Eval`) and fairness (`Fair Eval`), and ability to calibrate their predictions for the risk assessment model accuracy (`Acc Cal`) and fairness (`Fair Cal`). ✓indicates the desired behavior was observed (positive association, $p < 0.05$), and 0 indicates no statistically significant relationship.**

| | Conf | Acc Eval | Acc Cal | Fair Eval | Fair Cal |
|---|---|---|---|---|---|
| RA Only (unexplained) | 0 | 0 | ✓ | 0 | 0 |
| Diverse Counterfactual | 0 | ✓ | ✓ | 0 | 0 |
| Selective Counterfactual | 0 | 0 | ✓ | 0 | 0 |
| Complete Feature Attr. | 0 | 0 | ✓ | ✓ | 0 |
| Selective Feature Attr. | 0 | 0 | ✓ | 0 | 0 |

`Conf`: Within each treatment, we regressed self-reported performance against actual performance controlling for participant demographic information and exit survey responses. There were no statistically significant associations between self-reported confidence and actual performance in any treatment. Similarly, when we regressed self-reported confidence in performance relative to other participants (calculated as participant *prediction score* rank in the 559 worker sample) against actual performance, we found no statistically significant associations. This indicates that participants were unable to accurately assess their own performance neither absolutely, nor with respect to other non-expert participants.

To examine participants' evaluations of risk assessment model performance, we looked at two aspects of the model: accuracy and fairness. `Acc Eval`: Within each treatment, we regressed participants' self-reported perception of risk assessment model accuracy against the actual model accuracy they experienced controlling for participant performance, demographic information, and exit survey responses. Only the diverse counterfactual treatment showed a positive and statistically significant association ($p < 0.05$); participants were only able to successfully assess the accuracy of the risk assessment model in that treatment.

`Fair Eval`: As for assessing fairness, we similarly regressed within each treatment participants' self-reported perception of risk assessment model fairness against the actual model fairness they experienced (measured as difference in risk assessment model false positive rates for black and white defendants — the lower the difference, the higher the fairness) [3]. We only found a negative and statistically significant relationship between self-reported model fairness and the model's difference in false positive rates between black and white defendants in the complete feature attribution treatment; participants were only able to successfully assess risk assessment model fairness in that treatment.

*6.5.2 Calibration.* To look at whether participants adjusted their use of the risk assessment model according to the model's performance, we regressed within each treatment the influence of the model on each participant against the model (1) accuracy, and (2) fairness they experienced[4]. We found positive and statistically significant associations between influence and risk assessment model accuracy across all treatments ($p < 0.008$) (`Acc Cal`), but no statistically significant associations between influence and risk assessment model fairness (`Fair Cal`). Participants in the explanation treatments were able to adjust the influence of the model on their predictions according to its accuracy but not its fairness.

*6.5.3 Accountability.* To assess whether participants sensibly adjusted how much accountability they should face compared to the algorithm developers, we utilized their response to this survey question, "If one of the decisions you make goes wrong or is questioned, how much accountability do you think you should face?" Within each treatment, we regressed their response against the overall influence of the risk assessment model on their predictions controlling for risk assessment model performance, participant performance, demographic information, and exit survey responses. The diverse counterfactual treatment showed a positive and statistically significant correlation between influence and how much participants think they should be held accountable ($p < 0.003$). However, the degree of accountability was also associated with their self-reported confidence ($p < 0.05$), and how fair they perceived the algorithm to be: the fairer the algorithm, the more participants believed they should be held accountable ($p < 0.001$). As for the other treatments, the unexplained risk assessment model treatment and the complete feature attribution treatments showed a positive and statistically significant relationship between participant self-reported confidence and accountability. The more confident participants were in their performance, the more they believed they should be held accountable ($p < 0.05$).

## 6.6 Principle 3: Fairness

*6.6.1 Influence Disparity.* We measured influence disparity for cases where the risk assessment model prediction was greater than the average baseline prediction ($RA > baseline$) and for cases where it was less than the average baseline prediction ($RA < baseline$). For both $RA < baseline$ and the $RA > baseline$ cases, there were no statistically significant results.

*6.6.2 Deviation Disparity.* There was a statistically significant difference between the average deviation for black and white defendants across all treatments ($p < 0.04$). The average black deviation was more negative than the average white deviation in every treatment; participants on average deviated to scores lower than the model scores for black defendants compared to white defendants. This difference in average deviation was greater in the explanation treatments than in the unexplained risk assessment model treatment, where the largest difference of 0.59 was observed in the diverse counterfactual treatment, and the smallest of 0.15 was observed in the unexplained risk assessment model treatment.

## 6.7 Principle 4: Explanation Effectiveness

*6.7.1 Trust.* A prerequisite of trust is the proper evaluation of the risk assessment model's accuracy and fairness, in order to calibrate trust accordingly. As seen in Section 6.5, participants largely failed to properly assess the model. Thus, the appropriate calibration of trust was challenging to measure.

*6.7.2 Insight.* We first looked at how valuable and useful participants reported the explanations were to their decision making process. The average response to the survey question on how useful the explanations were was between "Moderately" (3) and "Very" (4) (average of 3.21) useful for all treatments except the selective counterfactual treatment, which was between "Slightly" (2) and "Moderately" (3) (average of 2.85) useful. This indicates that participants found the explanations to be reasonably informative.

We then examined whether the information gained from explanations led to improved participant performance and an improved participant ability to articulate how they arrived at the joint prediction outcome. First, within each treatment, we regressed participant performance against their self-reported degree to which they found the explanation useful controlling for risk assessment model performance, participant demographic information, and exit survey responses. We found no statistically significant associations between participant performance and self-reported explanation usefulness in any treatment; the increased informativeness of explanations did not lead to improved performance.

We then regressed within each treatment participant self-reported ability to explain how they arrived at their predictions against their self-reported degree to which they found the explanation useful controlling for risk assessment model performance, participant performance, demographic information, and exit survey responses. We only found a positive and statistically significant ($p < 0.05$) relationship between participants' self-reported explanation usefulness and ability to explain their decision making process in the selective feature attribution treatment.

*6.7.3 Fair & Ethical Decision-Making.* We first examined whether there was a reduction in decision-maker 'bias' in the explanation treatments, where 'bias' was defined in a similar manner to model fairness: difference in the participant false positive prediction rates for black and white defendants [11]. No explanation treatment led to a statistically significant decrease in participant bias relative to the unexplained risk assessment model treatment.

We then investigated whether explanations improved performance in the presence of false positive and false negative risk

[3] To focus on the most prevalent aspect of bias, this analysis was restricted to the 72% of participants who experienced a greater or equal false positive rate for black than white defendants.

[4] Most participants (83%) had influence values between 0 and 1. Those who adjusted beyond the risk assessment model or went against the model did not and were excluded from this analysis.

assessment model predictions. We analyzed a subset of the data where the model score presented was a false positive or false negative. We used a linear mixed-effects model with random effects for participant and defendant identities, and regressed participant performance against participant treatment controlling for risk assessment model performance, participant demographic information, and the defendant features. Only the diverse counterfactual treatment case led to a statistically significant ($p < 0.05$) increase in participant performance, relative to the unexplained risk assessment model treatment, while the defendant features had no statistically significant effect on performance.

## 7 DISCUSSION & FUTURE WORK

*7.0.1 Explanation Informativeness & Human Performance.* Despite participants reporting that explanations were reasonably informative, the explanation treatments did not improve accuracy relative to the unexplained risk assessment model treatment. This counterintuitive result is opposite to common assumptions that explanations improves performance (via improved human interpretability of algorithms) [35, 40], but is in line with more recent empirical findings on explanations and human performance [39, 45].

Furthermore, explanations did not lead to an increase in participants' self-reported confidence in articulating how they arrived at the joint prediction outcome. Thus, even if explanations had improved human performance, this does not necessarily indicate that explanations were “interpretable" to the degree that is necessary to ensure decision-makers feel they can adequately explain the joint prediction outcome — a critical element for effective accountability measures [10].

*7.0.2 Challenges of Assessing & Addressing Fairness.* In only one treatment did the risk assessment model exert more influence at increasing risk score of black defendants compared to white defendants. Apart from this, we observed less disparate interactions than we expected (and compared to [16]); our analysis showed that participants on average reduced the model's scores of black defendants more than white defendants. Survey responses provide evidence that participants had a preconceived awareness of algorithmic bias against black defendants and adjusted their predictions accordingly — possibly a result of increased activism and scholarship on racial bias in algorithms, especially in criminal justice [7, 37]. In terms of human-algorithm interaction, this increased awareness raises questions on the nature and extent of human adjustments to perceived bias in algorithms [26, 43]. Our results showed that participants in most of the treatments still fell short of accurately assessing risk assessment model fairness, calibrating the influence of the model scores, and correcting for false positive and false negative model predictions.

*7.0.3 The Argumentative Potential of Explanations.* Relative to when the risk assessment model was not explained, the diverse counterfactual was the only treatment shown to improve participants' performance in the presence of false positive and false negative model predictions. This result indicates that certain types of explanations could legitimize or disguise inaccurate and unfair AI predictions [13], motivating further empirical studies of xAI.

*7.0.4 DiCE User Feedback on Proximity & Diversity of Explanation Type.* A point of user feedback identified the need at times for factual explanations to accompany or replace the counterfactual ones. Our results lend some support to this feedback since we observed that factual and counterfactual explanations satisfied different criteria (e.g. assessing model fairness versus accuracy, respectively); rarely did both explanation types lead to the same desirable behavior. Indeed, prior work has also shown that users' preferences for counterfactual versus factual explanations vary according to the nature of each prediction and users' expectations of each prediction [41]. Thus, explanations that adapt to these varying needs, by combining factual and counterfactual explanations, may be more effective than one or the other.

This feedback, along with our observation that participants preferred and performed better in the diverse/complete than the selective treatments, reinstates the importance of studying the trade-off between complete or diverse explanations and the potential of information overload [20]. An important future axis of exploration can focus on what *Wachter et al.* identified as “social" or “interactive" explanations that involve iteration until the explanation leads to a point of mutual understanding between the explainer and explainee [32].

# A APPENDIX

## A.1 Web Experiment Task Interface

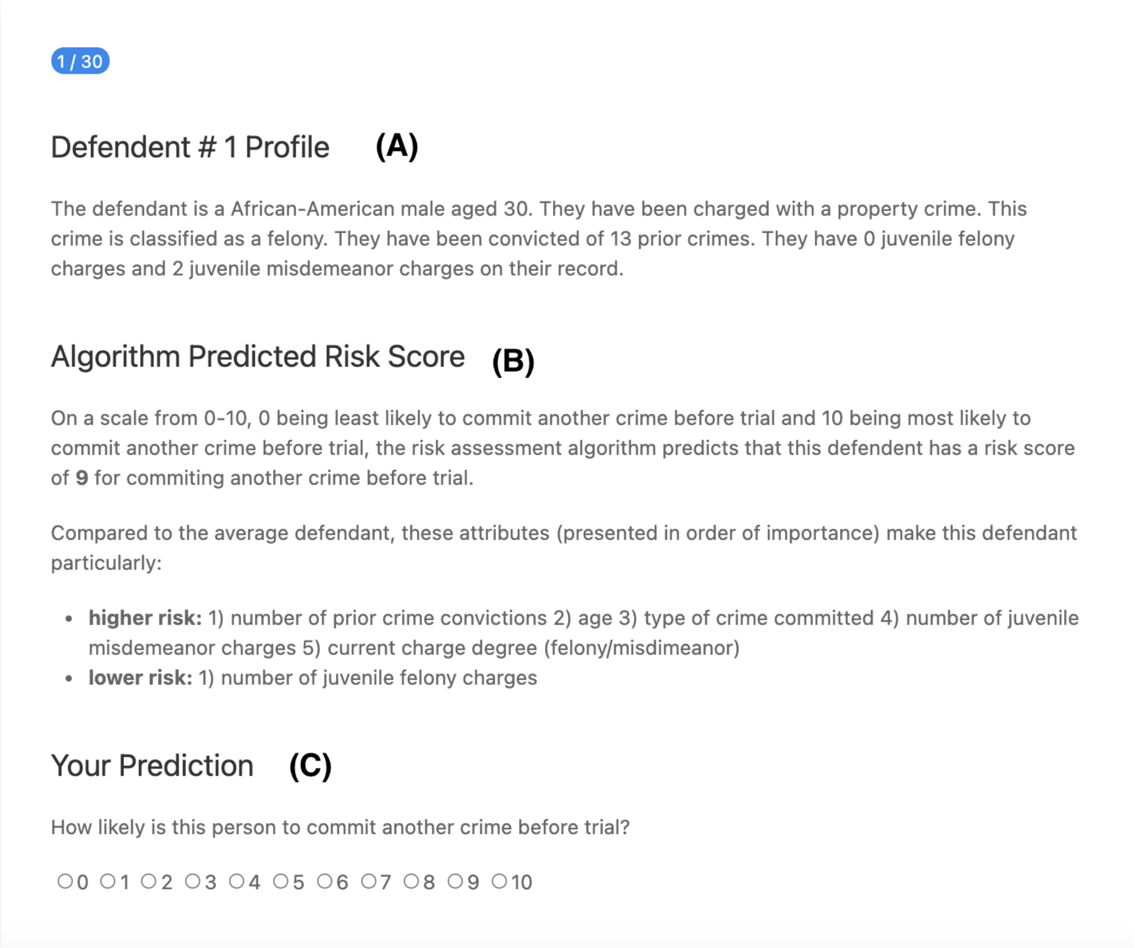
We provide an image in Figure 1 of the interface used to present tasks in the web-based experiemnt.

## A.2 Participant Qualitative Response on Explanations

Participants' responses to how they incorporated the explanations could be categorized into the following attitudes: (1) explanations were useful/used in their decision making process, (2) explanations were insightful to finding patterns in how the algorithm works, (3) explanations were not useful due to specific critiques, and (4) explanations were ignored for no specified reason.

For (1) and (2), participants mainly reported finding explanations insightful in these ways:

- Influencing their perceptions of what features were important for the prediction task (e.g., “I liked being able to see why something was high risk/versus lower risk. On a few instances, age was mentioned and I kept forgetting to take into account age so that affected my score.")
- Highlighting instances where the explanations/algorithm's prediction was erroneous/unfair (e.g., “I felt the algorithm was off base. It would state it would give a lower score if the defendant had more juvenile misdemeanor charges than he really had.", “... I also think there were similar circumstances where the only difference was race, yet the risk factor was judged quite differently in a few examples."
- Generating more curiosity about how the explanations relate to the risk score (e.g., “... I did wonder if that is what the algorithm's basis for scoring is, if the high and low risk variables even out, would they give a score of 5?")



1 / 30

Defendant # 1 Profile (A)

The defendant is a African-American male aged 30. They have been charged with a property crime. This crime is classified as a felony. They have been convicted of 13 prior crimes. They have 0 juvenile felony charges and 2 juvenile misdemeanor charges on their record.

Algorithm Predicted Risk Score (B)

On a scale from 0-10, 0 being least likely to commit another crime before trial and 10 being most likely to commit another crime before trial, the risk assessment algorithm predicts that this defendent has a risk score of **9** for commiting another crime before trial.

Compared to the average defendant, these attributes (presented in order of importance) make this defendant particularly:

- **higher risk:** 1) number of prior crime convictions 2) age 3) type of crime committed 4) number of juvenile misdemeanor charges 5) current charge degree (felony/misdimeanor)
- **lower risk:** 1) number of juvenile felony charges

Your Prediction (C)

How likely is this person to commit another crime before trial?

○0 ○1 ○2 ○3 ○4 ○5 ○6 ○7 ○8 ○9 ○10

**Figure 1: *Interface of a single task in the web-based experiment.* Tasks consisted of up to three sections: (A) defendant profile, (B) risk assessment model score (and accompanying explanation), and (C) participant prediction.**

- Creating more room for ‘human forgiveness’ (the absence of which is often cited in debates on human vs machine decision making [33]) in human-algorithm interactions (e.g., “If the person came close to the cutoff age that the AI said it [risk score] would have been lower, then I usually did lower the prediction number to correspond.")

As for critiques - (3) & (4) -, participants offered more, as well as more specific critiques for the counterfactual explanations. For the feature attribution explanations, critiques mainly referenced instances where participants did not agree with the high/low risk feature categorization. The two main critiques of the counterfactual explanations were the following:

- Explanations were irrelevant/factual explanations would have have been more relevant (e.g., “I did not [incorporate the explanations] because I did not think that they were very relevant. They told me how the algorithm could have decided differently but not why it decided the way it did."
- Explanations were not proximate enough to each defendant’s profile (e.g., “None of them seemed really appropriate so I started ignoring them - really high numbers of other convictions or weird crimes/unrelated or extreme age differences."

Critiques of the selective counterfactual explanations specifically, focused on disagreements with the explanations’ emphasis on age, despite age being regularly mentioned by participants as an important factor.

## REFERENCES

[1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[2] APPR. 2020. Public Safety Assessment: How It Works. https://advancingpretrial.org/psa/factors/#nca

[3] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (2020), 82–115.

[4] Mara S Aruguete, Ho Huynh, Blaine L Browne, Bethany Jurs, Emilia Flint, and Lynn E McCutcheon. 2019. How serious is the ‘carelessness’ problem on Mechanical Turk? *International Journal of Social Research Methodology* 22, 5 (2019), 441–449.

[5] Alejandro Barredo-Arrieta and Javier Del Ser. 2020. Plausible counterfactuals: Auditing deep learning classifiers with realistic adversarial examples. In *2020*

*International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–7.

[6] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (2020), 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

[7] Haydn Belfield. 2020. Activism by the AI community: Analysing recent achievements and future prospects. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 15–21.

[8] Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. 2015. The role of explanations on trust and reliance in clinical decision support systems. In *2015 international conference on healthcare informatics*. IEEE, 160–169.

[9] Ethan Corey. 2020. New Data Suggests Risk Assessment Tools Have Little Impact on Pretrial Incarceration. https://theappeal.org/new-data-suggests-risk-assessment-tools-have-little-impact-on-pretrial-incarceration/

[10] Finale Doshi-Velez, Mason Kortz, Ryan Budish, Chris Bavitz, Sam Gershman, David O'Brien, Stuart Schieber, James Waldo, David Weinberger, and Alexandra Wood. 2017. Accountability of AI Under the Law: The Role of Explanation. *SSRN Electronic Journal* (11 2017). https://doi.org/10.2139/ssrn.3064761

[11] Julia Dressel and Hany Farid. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science advances* 4, 1 (2018), eaao5580.

[12] Laurel Eckhouse, Kristian Lum, Cynthia Conti-Cook, and Julie Ciccolini. 2019. Layers of bias: A unified approach for understanding problems with risk assessment. *Criminal Justice and Behavior* 46, 2 (2019), 185–209.

[13] John Fox, David Glasspool, Dan Grecu, Sanjay Modgil, Matthew South, and Vivek Patkar. 2007. Argumentation-based inference and decision making–A medical perspective. *IEEE intelligent systems* 22, 6 (2007), 34–41.

[14] Jerome H Friedman. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics* (2001), 1189–1232.

[15] Ben Green. 2020. The false promise of risk assessments: epistemic reform and the limits of fairness. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 594–606.

[16] Ben Green and Yiling Chen. 2019. Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In *Proceedings of the conference on fairness, accountability, and transparency*. 90–99.

[17] Ben Green and Yiling Chen. 2019. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–24.

[18] Ben Green and Yiling Chen. 2020. Algorithm-in-the-loop decision making. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 13663–13664.

[19] Philipp Hacker, Ralf Krestel, Stefan Grundmann, and Felix Naumann. 2020. Explainable AI under contract and tort law: legal incentives and technical challenges. *Artificial Intelligence and Law* (2020), 1–25.

[20] Denis J Hilton. 1996. Mental models and causal explanation: Judgements of probable cause and explanatory relevance. *Thinking & Reasoning* 2, 4 (1996), 273–308.

[21] Nicholas C Hunt and Andrea M Scheetz. 2019. Using MTurk to distribute a survey or experiment: Methodological considerations. *Journal of Information Systems* 33, 1 (2019), 43–65.

[22] Hripsime A Kalaian and Stephen W Raudenbush. 1996. A multivariate mixed linear model for meta-analysis. *Psychological methods* 1, 3 (1996), 227.

[23] Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. 2013. Too much, too little, or just right? Ways explanations impact end users' mental models. In *2013 IEEE Symposium on visual languages and human centric computing*. IEEE, 3–10.

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

[25] Jeff Larson, Julia Angwin, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias. Retrieved Mar 1, 2021 from http://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[26] P. Law, Sana Malik, Fan Du, and M. Sinha. 2020. The Impact of Presentation Style on Human-In-The-Loop Detection of Algorithmic Bias. In *Graphics Interface*.

[27] Q Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: informing design practices for explainable AI user experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–15.

[28] Zachary C Lipton. 2018. The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* 16, 3 (2018), 31–57.

[29] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems*. 4768–4777.

[30] P Baumgartner MDeMichele, M Wenger, K Barrick, M Comfort, and S Misra. 2018. The Public Safety Assessment: A Re-Validation and Assessment of Predictive Utility and Differential Prediction by Race and Gender in Kentucky.(2018).

[31] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence* 267 (2019), 1–38.

[32] Brent Mittelstadt, Chris Russell, and Sandra Wachter. 2019. Explaining explanations in AI. In *Proceedings of the conference on fairness, accountability, and transparency*. 279–288.

[33] Carlos Montemayor, Jodi Halpern, and Abrol Fairweather. 2021. In principle obstacles for empathic AI: why we can't replace human empathy in healthcare. *Ai & Society* (2021), 1–7.

[34] Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. 2020. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 607–617.

[35] Menaka Narayanan, Emily Chen, Jeffrey He, Been Kim, Sam Gershman, and Finale Doshi velez. 2018. How do Humans Understand Explanations from Machine Learning Systems? An Evaluation of the Human-Interpretability of Explanation. (02 2018).

[36] Northpointe. 2010. http://www.northpointeinc.com/files/technical_documents/Selected_Compas_Questions_Posed_by_Inquiring_Agencies.pdf

[37] Leila Ouchchy, Allen Coin, and Veljko Dubljević. 2020. AI in the headlines: the portrayal of the ethical issues of artificial intelligence in the media. *AI & SOCIETY* 35, 4 (2020), 927–936.

[38] Wolter Pieters. 2011. Explanation and trust: what to tell the user in security and AI? *Ethics and information technology* 13, 1 (2011), 53–64.

[39] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–52.

[40] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.

[41] Maria Riveiro and Serge Thill. 2021. "That's (not) the output I expected!" On the role of end user expectations in creating explanations of AI systems. *Artificial Intelligence* 298 (2021), 103507. https://doi.org/10.1016/j.artint.2021.103507

[42] James Schaffer, John O'Donovan, James Michaelis, Adrienne Raglin, and Tobias Höllerer. 2019. I can do better than your AI: expertise and explanations. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 240–251.

[43] Philipp Schmidt and Felix Biessmann. 2020. Calibrating human-ai collaboration: Impact of risk, ambiguity and transparency on algorithmic bias. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*. Springer, 431–449.

[44] Ed Yong. 2018. A popular algorithm is no better at predicting crimes than random people. *The Atlantic* 17 (2018), 2018.

[45] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 295–305.