Mathematical Notions vs. Human Perception of Fairness: A Descriptive Approach to Fairness for Machine Learning

Megha Srivastava
Stanford University
meghas@stanford.edu

Hoda Heidari
ETH Zürich
hheidari@inf.ethz.ch

Andreas Krause
ETH Zürich
krausea@ethz.ch

ABSTRACT

Fairness for Machine Learning has received considerable attention, recently. Various mathematical formulations of fairness have been proposed, and it has been shown that it is impossible to satisfy all of them simultaneously. The literature so far has dealt with these impossibility results by quantifying the tradeoffs between different formulations of fairness. Our work takes a different perspective on this issue. Rather than requiring all notions of fairness to (partially) hold at the same time, we ask which one of them is the most appropriate given the societal domain in which the decision-making model is to be deployed. We take a descriptive approach and set out to identify the notion of fairness that best captures *lay people's perception of fairness*. We run adaptive experiments designed to pinpoint the most compatible notion of fairness with each participant's choices through a small number of tests. Perhaps surprisingly, we find that the most simplistic mathematical definition of fairness—namely, demographic parity—most closely matches people's idea of fairness in two distinct application scenarios. This conclusion remains intact even when we explicitly tell the participants about the alternative, more complicated definitions of fairness, and we reduce the cognitive burden of evaluating those notions for them. Our findings have important implications for the Fair ML literature and the discourse on formalizing algorithmic fairness.

ACM Reference Format:
Megha Srivastava, Hoda Heidari, and Andreas Krause. 2019. Mathematical Notions vs. Human Perception of Fairness: A Descriptive Approach to Fairness for Machine Learning. In *The 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '19), August 4–8, 2019, Anchorage, AK, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3292500.3330664

1 INTRODUCTION

Machine Learning tools are increasingly employed to make consequential decisions for human subjects, in areas such as credit lending [23], policing [25], criminal justice [3], and medicine [8]. Decisions made by these algorithms can have a long-lasting impact on people's lives and may affect certain individuals or social groups negatively [1, 28]. This realization has recently spawned an active area of research into quantifying and guaranteeing fairness for machine learning [9, 14, 19].

Despite the recent surge of interest in Fair ML, there is no consensus on a precise definition of (un)fairness. Numerous mathematical definitions of fairness have been proposed; examples include demographic parity [9], disparate impact [32], equality of odds [10, 14], and calibration [19]. While each of these notions is appealing on its own, it has been shown that they are incompatible with one another and cannot hold simultaneously [6, 19]. The literature so far has dealt with these impossibility results by attempting to quantify the tradeoffs between different formulations of fairness, hoping that practitioners will be better positioned to determine the extent to which the violation of each fairness criterion can be socially tolerated (see, e.g., [7]).

Our work takes a different perspective on these impossibility results. We posit that fairness is a highly context-dependent ideal and depending on the societal domain in which the decision-making model is deployed, one mathematical notion of fairness may be considered ethically more desirable than other alternatives. So rather than requiring all notions of fairness to (partially) hold at the same time, we set out to determine the most suitable notion of fairness for the particular domain at hand. Because algorithmic predictions ultimately impact people's lives, we argue that the most appropriate notion of algorithmic fairness is the one that reflects people's idea of fairness in the given context. We, therefore, take a *descriptive ethics* approach to **identify the mathematical notion of fairness that most closely matches lay people's perception of fairness**.

Our primary goal is to test the following hypotheses:[1]

- **H1:** In the context of recidivism risk assessment, the majority of subjects' responses is compatible with equality of *false negative/positive rates* across demographic groups.
- **H2:** In the context of medical predictions, the majority of subjects' responses is compatible with *equality of accuracy* across demographic groups.
- **H3:** When the decision-making stakes are high (e.g., when algorithmic predictions affect people's life expectancy) participants are more sensitive to accuracy as opposed to equality.

Our first hypothesis is inspired by the recent media coverage of the COMPAS criminal risk assessment tool and its potential bias against African-American defendants [1]. Our second hypothesis is informed by the growing concern over the fact that medical experiments are largely conducted with white males as subjects and as a result, their conclusions are reliable only for that particular segment of the population [29]. Such biases can only be amplified

KDD '19, August 4–8, 2019, Anchorage, AK, USA

ACM ISBN 978-1-4503-6201-6/19/08…$15.00
https://doi.org/10.1145/3292500.3330664

[1] All hypotheses were defined and registered with our institution in advance of the experiments.

by Machine Learning tools [15]. Our third hypothesis explores the tension between inequality aversion and accuracy.

In order to determine the notion of fairness that is most compatible with a participant's perception of fairness, we design an adaptive experiment based on active learning. Each participant is required to answer a series of at most 20 adaptively chosen tests—all concerning a fixed, carefully specified scenario (e.g., predicting the risk of future crime for defendants). Each question specifies the ground truth labels for ten hypothetical decision subjects, along with the labels predicted for them by two hypothetical predictive models./algorithms (see Figure 1). The participant is then prompted to choose which of the two algorithms they consider to be more discriminatory. The tests are chosen by an active learning algorithm, called EC^2 [12, 24]. Depending on the participant's choices so far, the EC^2 algorithm chooses the next test. Note that an active learning scheme is necessary for pinpointing the most compatible notion through a small number of tests. Without an adaptive design, participants would have to answer hundreds of questions before we can confidently specify the notion of fairness that is most compatible with their choices.

We run our experiments on the Amazon Mechanical Turk (AMT) platform and report the percentage of participants whose choices match each mathematical notion of fairness. We also investigate how this varies from one context to another and from one demographic group to another. Surprisingly, we find that the most simplistic mathematical definition of fairness—namely, *demographic parity*—most closely matches people's idea of fairness in two distinct scenarios. This remains the case even when subjects are explicitly told about the alternative, more complicated definitions of fairness and the cognitive burden of evaluating those notions is reduced for them. Our findings have important implications for the Fair ML literature. In particular, they accentuate the need for a more comprehensive understanding of the human attitude toward fairness for predictive models.

In summary, we provide a comparative answer to the ethical question of "what is the most appropriate notion of fairness to guarantee in a given societal context?". Our framework can be readily adapted to scenarios beyond the ones studied here. As the Fair ML literature continues to grow, we believe it is critical for all stakeholders—including people who are potential subjects of algorithmic decision-making in their daily lives—to be heard and involved in the process of formulating fairness. Our work takes an initial step toward this agenda. After all, a theory of algorithmic fairness can only have a meaningful positive impact on society if it reflects people's sense of justice.

1.1 Related Work

Much of the existing work on algorithmic fairness has been devoted to the study of *discrimination* (also called *statistical-* or *group*-level fairness) for *binary classification*. Statistical notions require a particular metric—quantifying *benefit* or harm—to be equal across different social groups (e.g., gender or racial groups). Different choices for the benefit metric have led to different fairness criteria; examples include Demographic Parity (DP) [9], Error Parity (EP) [5], equality of False Positive or False Negative rates (FPP and FNP, respectively) [14, 32], and False Discovery or Omission rates (FDP and FOP, respectively) [19, 32]. Demographic parity seeks to equalize the percentage of people who are predicted to be positive across different groups. Equality of false positive/negative rates requires the percentage of people falsely predicted to be positive/negative to be the same across true negative/positive individuals belonging to each group. Equality of false discovery/omission rates seeks to equalize the percentage of false positive/negative predictions among individuals predicted to be positive/negative in each group. See Table 1 for the precise definition of benefit corresponding to each notion of these notions of fairness.

Table 1: The measure of benefit/harm corresponding to each notion of fairness. For a decision subject i belonging to group G, y_i specifies his/her true label, and $\hat{y}_i$ his/her predicted label. n_G is the number of individuals belonging to group G.

Fairness notion	benefit for group G
DP	$b^G = \frac{1}{n_G}\sum_{i\in G} 1[\hat{y}_i = 1]$
EP	$b^G = \frac{1}{n_G}\sum_{i\in G} 1[\hat{y}_i \neq y_i]$
FDP	$b^G = \frac{\sum_{i\in G} 1[y_i=0 \& \hat{y}_i=1]}{\sum_{i\in G} 1[\hat{y}_i=1]}$
FNP	$b^G = \frac{\sum_{i\in G} 1[\hat{y}_i=0 \& y_i=1]}{\sum_{i\in G} 1[y_i=1]}$

Following Speicher et al. [27], we extend these existing notions of fairness to measures of unfairness by employing inequality indices, and in particular the Generalized Entropy index. Given the benefit vector, $\mathbf{b}$, computed for all groups $G = 1, \cdots, N$, the Generalized Entropy (with exponent $\alpha = 2$) is calculated as follows:

$$\mathcal{E}(\mathbf{b}) = \mathcal{E}(b_1, b_2, \cdots, b_N) = \frac{1}{2n}\sum_{G=1}^{N}\left[\left(\frac{b_G}{\mu}\right)^2 - 1\right]. \qquad (1)$$

where $\mu = \frac{1}{N}\sum_{G=1}^{N} b_G$ is the average benefit across all groups.

In ethics, there are two distinct ways of addressing moral dilemmas: *descriptive* vs. *normative* approach. Normative ethics involves creating or evaluating moral standards to decide what people *should do* or whether their current moral behavior is reasonable. Descriptive (or comparative) ethics is a form of empirical research into the attitudes of individuals or groups of people towards morality and moral decision-making. Our work belongs to the descriptive category. Several prior papers have taken a normative perspective on algorithmic fairness. For instance, Gajane and Pechenizkiy [11] attempt to cast algorithmic notions of fairness as instances of existing theories of justice. Heidari et al. [16] propose a framework for evaluating the assumptions underlying different notions of fairness by casting them as special cases of economic models of equality of opportunity. We emphasize that there is no simple, widely-accepted, normative principle to settle the ethical problem of algorithmic fairness.

Several recent papers empirically investigate the issues of fairness and interpretability utilizing human-subject experiments. Below we elaborate on some of the prior work in this line. MIT's moral machine [2] provides a crowd-sourcing platform for aggregating human opinion on how self-driving cars should make decisions when faced with moral dilemmas. For the same setting, Noothigattu et al.

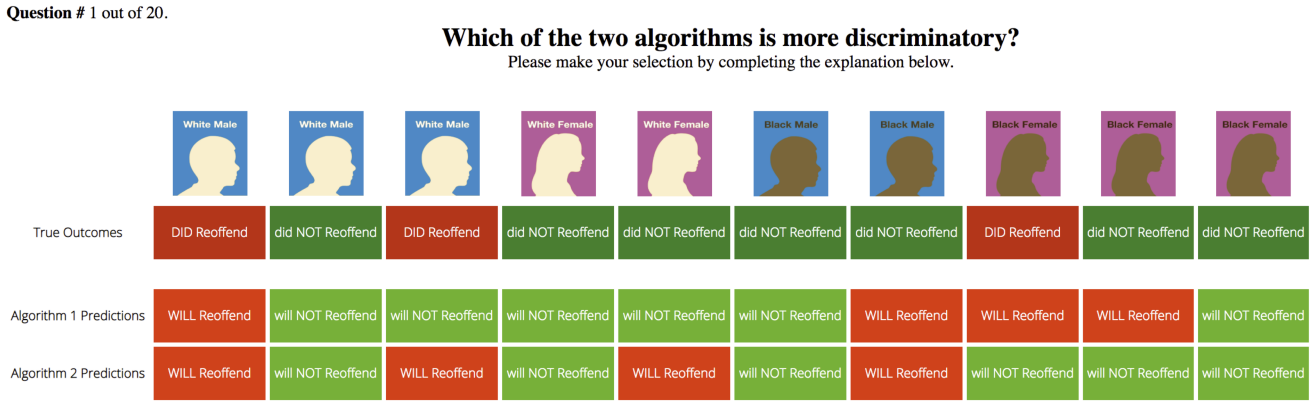

Figure 1: A typical test in our experiments. Each test illustrates the predictions made by two hypothetical algorithms (A_1 and A_2) along with the true labels for ten hypothetical individual decision subjects. Decision subjects' race and gender are color-coded. A pink/blue background specifies a female/male decision-subject. A beige/brown contour specifies a Caucasian/African-American decision subject. The demographic characteristics of decision subjects are kept constant across all tests. With this information, the participant responds to the test by selecting the algorithm he/she considers to be discriminatory.

[22] propose learning a random utility model of individual preferences, then efficiently aggregating those individual preferences through a social choice function. [21] proposes a similar approach for general ethical decision-making. Similar to these papers, we obtain input from human-participants by asking them to compare two alternatives from a moral standpoint. Noothigattu et al. and Lee et al. focus on modeling human preferences and *aggregating* them utilizing tools from *social choice theory*. In contrast, our primary goal is to understand the relationship between human perception and the recently proposed mathematical formulations of fairness.

Grgic-Hlaca et al. [13] study why people perceive the use of certain features as unfair in making a prediction about individuals. Binns et al. [4] study people's perceptions of justice in algorithmic decision-making under different *explanation styles* and at a high level show that there may be no "best" approach to explaining algorithmic decisions to people. Veale et al. [30] interview *public sector machine learning practitioners* regarding the challenges of incorporating public values into their work. Holstein et al. [17] conduct a systematic investigation of *commercial product teams*' challenges and needs in developing fairer ML systems through semi-structured interviews. Unlike our work where the primary focus is on lay people and potential decision subjects, Holstein et al. and Veale et al. study ML practitioners' views toward fairness.

Several recent papers in human-computer interaction study users' expectations and perceptions related to fairness of algorithms. Lee and Baykal [20] investigate people's perceptions of *fair division* algorithms (e.g., those designed to divide rent among tenants) compared to discussion-based group decision-making methods. Woodruff et al. [31] conduct workshops and interviews with participants belonging to certain marginalized groups (by race or class) in the US to understand their reactions to algorithmic unfairness.

To our knowledge, no prior work has conducted experiments with the goal of mapping existing *group* definitions of fairness to human perception of justice. Saxena et al. [26] investigate ordinary people's attitude toward three notions of *individual fairness* in the context of *loan decisions*. They investigate the following three notions: 1) treating similar individuals similarly [9]; 2) never favoring a worse individual over a better one [18]; 3) the probability of approval being proportional to the chance of the individual representing the best choice. They show that people exhibit a preference for the last fairness definition—which is similar, in essence, to calibration.

2 STUDY DESIGN AND METHODOLOGY

To test **H1** and **H2**, we conducted human-subject experiments on AMT to determine which one of the existing notions of group fairness best captures people's idea of fairness in the given societal context. For our final hypothesis, **H3**, we utilized short survey questions (see Section 4 for further details).

2.1 User Interfaces

In our experiments, each participant was asked to respond to a maximum of 20 tests. In each test, we showed the participant the predictions made by two hypothetical algorithms for ten hypothetical individuals, along with the true labels (see Figure 1). We limited attention to tests that consist of algorithms with equal overall accuracy to control for the impact of accuracy on people's perception. Note that the pair of algorithms displayed in each test changes from one test to another. We asked the participant to specify which algorithm they consider to be more discriminatory. The tests were chosen adaptively to find the notion of fairness most compatible with the participant's choices using only a small number of tests. The participant was then required to provide an *explanation* for their choice.

To elicit the explanations, we designed and tested two different User Interfaces (UI). The first UI presents the participant with

a text box, allowing them to provide an unstructured explanation/reasoning for their choice. The second UI requires the participant to provide a structured explanation, and it consists of two drop-down menus. In the first dropdown menu, the participant must choose the demographic characteristic they think the algorithm is most discriminatory with respect to (that is, gender, race, or the intersection of the two). In the second drop-down menu, they choose the metric along which the think the algorithm is most discriminatory. To reduce the cognitive burden, we computed the benefit metric corresponding to each notion of fairness and displayed it in the second drop-down menu. See Figure 2.

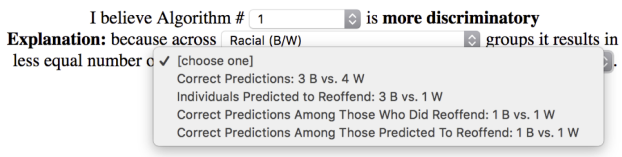

Figure 2: The user interface eliciting structured explanations from participants. All benefit metrics are computed and displayed to reduce the cognitive burden of evaluating our fairness notions.

We validated our interface design through two rounds of *pilot* studies—one internal (among members of our research group) and one on AMT (among 20 crowd workers). We made some minor changes after our internal run to improve readability of the task description. The AMT pilot participants found our first user interface (the one with text explanations) less restrictive and easier to work with, so we scaled up our experiments to 100 crowd workers per scenario using that interface.

2.2 Scenarios and Contexts

We ran our experiments for two distinct prediction tasks: criminal and skin cancer risk prediction. Below we describe each scenario precisely as it was shown to the study participants.

Criminal risk prediction. *Across the United States, data-driven decision-making algorithms are increasingly employed to predict the likelihood of future crimes by defendants. These algorithmic predictions are utilized by judges to make sentencing decisions for defendants (e.g., setting the bond amount; time to be spent in jail). Data-driven decision-making algorithms use historical data about past defendants to learn about factors that highly correlate with criminality. For instance, the algorithm may learn from past data that: 1) a defendant with a lengthy criminal history is more likely to reoffend if set free—compared to a first-time defender, or 2) defendants belonging to certain groups (e.g., residents of neighborhoods with high crime rate) are more likely to reoffend if set free. However, algorithms are not perfect, and they inevitably make errors—-although the error rate is usually very low, the algorithm's decision can have a significant impact on some defendants' lives. A defendant falsely predicted to reoffend can unjustly face longer sentences, while a defendant falsely predicted not to reoffend may commit a crime that was preventable.*

Skin cancer risk prediction. *Data-driven algorithms are increasingly employed to diagnose various medical conditions, such as risk for heart disease or various forms of cancer. They can find patterns and links in medical records that previously required great levels of expertise and time from human doctors. Algorithmic diagnoses are then used by health-care professionals to create personalized treatment plans for patients (e.g., whether the patient should undergo surgery or chemotherapy). Data-driven decision-making algorithms use historical data about past patients to learn about factors that highly correlate with risk of cancer. For instance, the algorithm may learn from past data that: 1) a patient with a family history of skin cancer has a higher risk of developing skin cancer; or 2) patients belonging to certain groups (e.g., people with a certain skin tone, or people of a certain gender) are more likely to develop skin cancer. However, algorithms are not perfect, and they inevitably make errors—although the error rate is usually very low, the algorithm's decision can have a significant impact on patients' lives. A patient falsely diagnosed with high risk of cancer may unnecessarily undergo high-risk and costly medical treatments, while a patient falsely labeled as low-risk for cancer may face a lower chance of survival.*

2.3 Adaptive Experimental Design

Each run of our experiment consists of sequentially selecting among a set of noisy tests and observing the participant's response to it. These tests are time-consuming for participants to evaluate and they can quickly become repetitive.[2] So to limit the number of tests per participant, we employed an active learning scheme proposed by Golovin et al. [12]. This allowed us to determine the most compatible fairness notion with each participant's choices, employing at most 20 tests per participant (see Figure 3). The algorithm, called EC^2 (for Equivalence Class Edge Cutting algorithm), can handle noisy observations and its expected cost is competitive with that of the optimal sequential policy.

The EC^2 algorithm. At a high level, the algorithm works as follows: it designates an *equivalence class* for each notion of fairness. In our setting, we have four equivalence classes representing Demographic Parity (DP), Error Parity (EP), False Discovery rate Parity (FDP), and False Negative rate Parity (FNP).[3] We use the notation h_1, h_2, h_3, h_4 to refer to these notions, respectively. We assume a uniform prior over $h_1, \cdots, h_4$.

Let τ denote the number of all possible tests we can run. In our case, $\tau = 9262$. We assume the cost is uniform across all tests. The outcome of each test is binary (the participant chooses one of the two algorithms as more discriminatory), so there will be 2^τ possible test outcome profiles, where a test outcome profile specifies the outcome of all τ tests. For each test outcome profile, EC^2 computes the fairness notion that maximizes the posterior probability of that outcome profile; we call this the MAP fairness notion for the outcome profile. EC^2 puts each outcome profile in the equivalence class corresponding to its MAP fairness notion.

EC^2 introduces edges between any two outcome profiles that belong to different equivalence classes. As the algorithm progresses, some edges are cut, and the algorithm terminates when no edge

[2] Every participant who completed our task was paid 5 USD.
[3] For additional experiments excluding DP and including other group notions of fairness, we refer the reader to the complete version of this paper on ArXiv.

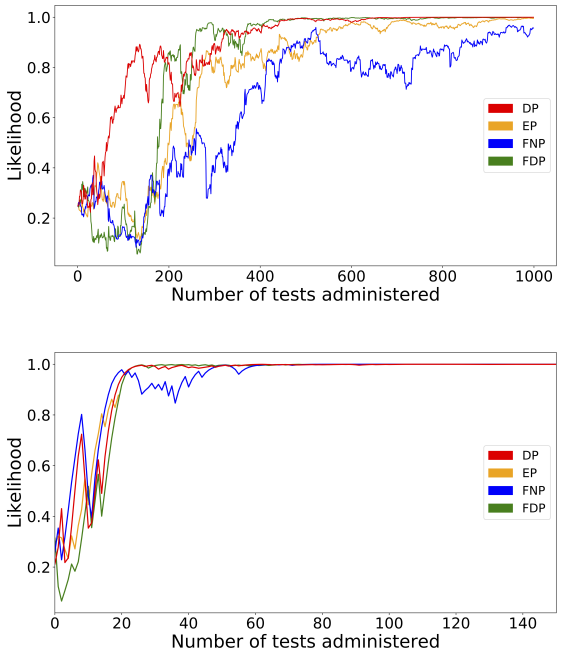

Figure 3: The advantage of adaptive over random test selection. For each of the four notions of fairness, we simulated a participant who follows that notion—as defined by our noisy response model—for 1000 tests. (Top) With a random test sequence, at least 600 tests are needed before we can obtain a high likelihood for one hypothesis. (Bottom) With our adaptive test selection method, only around 20 tests are needed to assign a high likelihood to the fairness notion the participant is following.

is left. At a high level, in each step, the algorithm picks the test that if run, in expectation removes as much of the remaining edge weights as possible.[4]

The Bayesian update. Let $\mathcal{T}$ denote the set of all available tests, and $\mathcal{A}$ the set of tests already carried out in the preceding steps. Let O_t = denote the outcome of the test currently administered, denoted by t (i.e., $O_t \in \{A_1, A_2\}$), and $\mathbf{o}$, the vector of outcomes of tests in $\mathcal{A}$. Then at each step we select the test $t \in \mathcal{T} - \mathcal{A}$ such that the following objective function is maximized:

$$\triangle(t|\mathbf{o}) = \left[\sum_{o\in\{A_1,A_2\}} \mathbb{P}(O_t = o|\mathbf{o})\left(\sum_i \mathbb{P}(h_i|\mathbf{o}, O_t = o)^2\right)\right] - \sum_i \mathbb{P}(h_i|\mathbf{o})^2$$

The objective function for each candidate test is computed as follows:[5] We assume a uniform prior on our four hypotheses, that is, for $h \in \{h_1 \cdots, h_4\}$, $\mathbb{P}(h) = 1/4$. We assume conditional independence, that is, $\mathbb{P}(\mathbf{o}|h) = \prod_{o\in\mathbf{o}} \mathbb{P}(o|h)$. We have that $\mathbb{P}(\mathbf{o}) = \sum_h \mathbb{P}(\mathbf{o}|h)\mathbb{P}(h)$ and $\mathbb{P}(h|\mathbf{o}) = \frac{\mathbb{P}(\mathbf{o}|h)\mathbb{P}(h)}{\mathbb{P}(\mathbf{o})}$. Moreover, $\mathbb{P}(O_t = o|\mathbf{o}) = \sum_h \mathbb{P}(O_t = o|h)\mathbb{P}(h|\mathbf{o})$. Calculating $\mathbb{P}(O_t = o|h)$ requires modeling participant's fairness assessments, which we discuss next.

Modeling participants' fairness assessments. After observing the participant's response to a test, EC^2 needs to update the posterior probability of each notion of fairness. This requires us to specify a model of how the participant responds to each test conditioned on their preferred notion of fairness. We assume[6] that the probability of an individual following h_i selecting A_1 is as follows:

$$\mathbb{P}(A_1|h_i) = \text{softmax}\left(\mathcal{E}(\mathbf{b}_{i,1}), \mathcal{E}(\mathbf{b}_{i,2})\right) \quad (2)$$

where $\mathbf{b}_{i,1} = \langle b^G_{i,1}\rangle_{\text{group } G}$ and $\mathbf{b}_{i,2} = \langle b^G_{i,2}\rangle_{\text{group } G}$ are the group-level benefit vectors for A_1 and A_2 respectively. $\mathcal{E}(\mathbf{b})$ is the Generalized Entropy index, a measure of inequality, computed over the benefit vector $\mathbf{b}$. For the benefit vector corresponding to each notion of fairness, see Table 1.

Figure 4 shows the outcome of our adaptive experiments, assuming that participants choose their answer to each test uniformly at random.

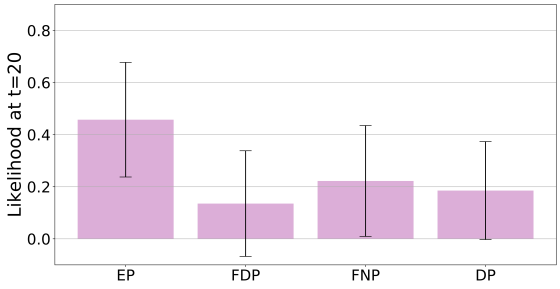

Figure 4: The outcome of our adaptive experiment if the response to each test is chosen at random. The chance of DP being selected as the most compatible hypothesis is not high.

3 EXPERIMENTAL FINDINGS

Through our experiments on AMT, we gathered a data set consisting of

- 100 participants' responses to our tests for the crime risk prediction scenario;
- 100 participants' responses to our tests for the skin cancer risk prediction scenario.

We asked the participants to provide us with their demographic information, such as their age, gender, race, education, and political affiliation. The purpose of asking these questions was to understand whether there are significant variations in perceptions of fairness across different demographic groups.[7] Table 2 summarizes the demographic information of our participants and contrasts it with the 2016 U.S. census data. In general, AMT workers are not a representative sample of the U.S. population (they often belong

[4] We implemented an efficient approximation of EC^2, called the EffECXtive (Efficient Edge Cutting approXimate objective) algorithm.

[5] The implementation makes use of memorization in order to speed up the computation and achieve a responsive user interface.

[6] There are numerous alternative assumptions we could have made; we consider ours to be one reasonable choice.

[7] Answering to this part was entirely optional and did not affect the participant's eligibility for participation or monetary compensation.

to a particular segment of the population—those that have Internet access and are willing to complete crowdsourced tasks online). In particular, for our experiments, participants were younger and more liberal compared to the average U.S. population.

Table 2: Demographics of our AMT participants compared to the 2016 U.S. census data.

Demographic Attribute	AMT	Census
Male	53%	49%
Female	47%	51%
Caucausian	68%	61%
African-American	12%	13%
Asian	10%	6%
Hispanic	6%	18%
Liberal	74%	33%
Conservative	19%	29%
High school	31%	40%
College degree	48%	48%
Graduate degree	20%	11%
18–25	14%	10%
25–40	67%	20%
40–60	16%	26%

3.1 Quantitative Results

For the crime risk prediction scenario, Figure 5 shows the number of participants whose responses were compatible with each notion of fairness, along with the confidence with which the EC^2 algorithm puts them in that category. Demographic parity best captures the choices made by the majority of our participants. Trends are similar for the cancer risk prediction scenario. See Table 3 for the summary.

Table 3: Number of participants matched with each notion with high likelihood ($> 80\%$).

	DP	EP	FDR	FOP	none
Crime risk prediction	80%	0%	2%	4%	14%
Cancer risk prediction	73%	3%	0%	0%	24%

Sensitivity to explanation interface. To test the sensitivity of our findings to the design of the user interface, we experimented (at a smaller scale with 20 participants) with an interface that elicits structured explanations from the participant. The interface explicitly shows the disparities across $h_1, \cdots, h_4$. Through this interface, we prompted the participant to think about the fairness notions of interest and reduced the cognitive burden of evaluating those notions for them. We observe that even under this new condition the majority of participants made choices that are best captured by demographic parity. For the crime prediction scenario, 17 out of 20 participants were matched with DP, and for the cancer prediction scenario, this number was 9 out of 20.

Table 4: Typical explanations provided by participants in each category.

Category	Explanation
DP	*"It expected only black females to reoffend."*
EP	*"Algorithm 1 made more correct decisions for white people."*
FDR	*"White males are incorrectly labeled likely to reoffend more often by this algorithm than the other."*
FNR	*"Two white males who did reoffend, both received* will not reoffend *status."*

Variation across gender, race, age, education, and political views. We did not observe significant variation across any demographic attribute. For the crime risk prediction scenario, the percentage of subjects whose likelihood of following DP is high ($> 80\%$) is as follows:

- 78% for female, and 79% for male participants;
- 82% for Caucasian, and 72% for non-Caucasian participants;
- 80% for liberal, and 74% for conservative participants;
- 77% for participants with no college/university degree, and 79% for the rest;
- 78% for young participants (with age < 40), and 81% for older participants.

3.2 Qualitative Results

Table 4 shows several instances of the explanations provided by the participants in each category. For example, a participant categorized by our system as following the Error Parity hypothesis provided an explanation that focused on the accuracy rate within demographic groups. These explanations demonstrate the ability for human participants to provide explanations in free text that can align with mathematical notions of fairness, without being prompted to think along these statistical definitions.

The free text explanations provided by participants present us with insight into their trajectories, and why a small number of participants could not be confidently categorized as either one of the four notions. Figure 7 compares the trajectories of two participants - one categorized as following Demographic Parity with high confidence, and another participant who is not confidently categorized into any metric. The latter participant's explanations vary across time steps, from considering false negative rates to accuracy within demographic groups. These inconsistent explanations support our system's inability to assign a fairness notion with high confidence.

4 SURVEY DESIGN AND ANALYSIS

To test **H3**, we presented participants with three algorithms, each offering a different trade-off between accuracy and equality. We asked our participants to choose the one they consider ethically more desirable. For the case of medical risk prediction, we performed this survey for two different scenarios: 1) predicting the risk of skin cancer (high-stakes), and predicting the severity of flu symptoms (low stakes). We hypothesized that when the stakes are high, more participants would choose high overall accuracy (i.e., A_1) over low inequality (i.e., A_3). Similarly, for the case of

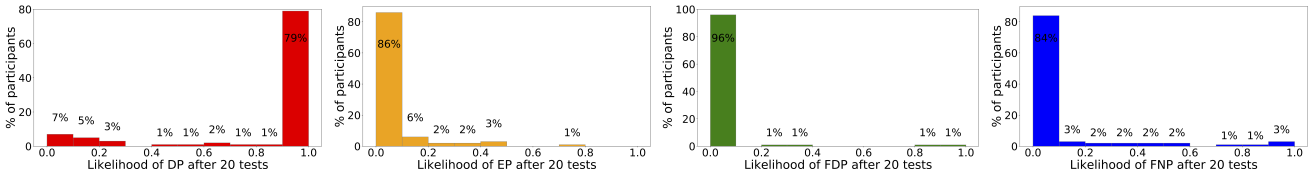

Figure 5: The crime risk prediction scenario—the number of participants matched with each notion of fairness (y-axis) along with the likelihood levels (x-axis). Demographic parity captures the choices made by the majority of participants.

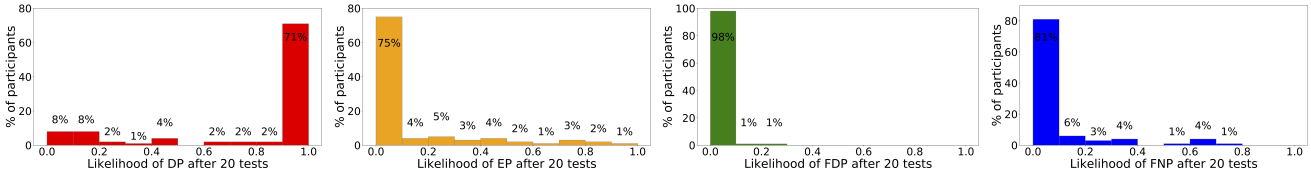

Figure 6: The cancer risk prediction scenario—the number of participants matched with each notion of fairness (y-axis) along with the likelihood levels (x-axis). Demographic parity captures the choices made by the majority of participants.

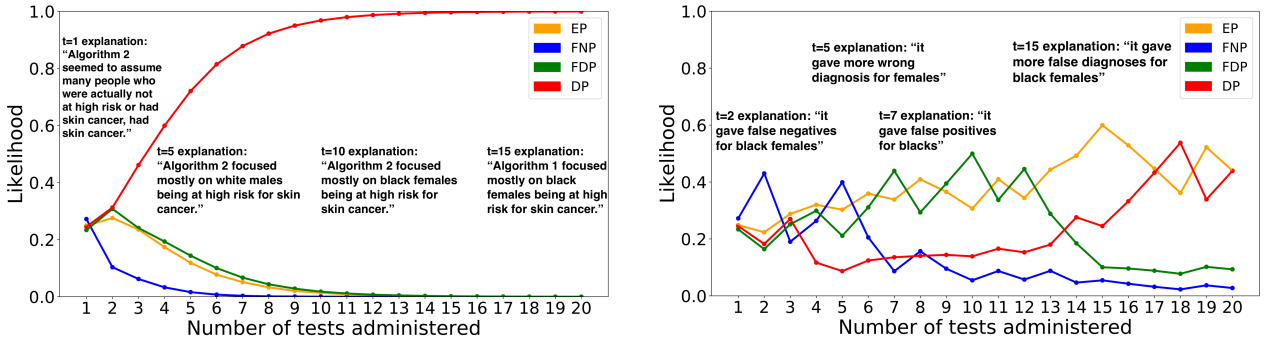

Figure 7: Trajectories of two different participants, one with a high final likelihood on DP and one with a low final likelihood on any fairness notion. Explanations provided by the former participant demonstrates consistency with an explanation aligned with demographic parity—addressing the predictions without considering ground truth labels. Explanations provided by the latter participant demonstrates inconsistency across different fairness notions.

criminal risk assessment, we conducted the survey for both a high-stakes (predictions used to determine jail time) and a low-stakes scenario (predictions used to set bail amount.) Below we lay out the questionnaires precisely as they were presented to the survey participants.

Table 5: Three hypothetical algorithms offering distinct tradeoffs between accuracy and inequality.

Algorithm	accuracy	female acc.	male acc.
A_1	94%	89%	99%
A_2	91%	90%	92%
A_3	86%	86%	86%

Skin cancer risk prediction. *Data-driven algorithms are increasingly employed to screen and predict the risk of various forms of diseases, such as skin cancer. They can find patterns and links in medical records that previously required great levels of expertise and time from human doctors. Algorithmic predictions are then utilized by health-care professionals to create the appropriate treatment plans for patients. Suppose we have three skin cancer risk prediction algorithms and would like to decide which one should be deployed for cancer screening of patients in a hospital. Each algorithm has a specific level of accuracy—where accuracy specifies the percentage of subjects for whom the algorithm makes a correct prediction. See Table 5. Note that in cases where the deployed algorithm makes an error, a patient's life can be severely impacted. A patient falsely predicted to have high risk of skin cancer may unnecessarily undergo high-risk and costly medical interventions, while a patient falsely labeled as low risk for skin cancer may face a significantly lower life expectancy. From a moral standpoint, which one of the following three algorithms do you think is more desirable for deployment in real-world hospitals?*

Flu symptom severity prediction. *Data-driven algorithms can be employed to screen and predict the risk of various forms of diseases, including common flu. They can find patterns and links in medical*

records that previously required visiting a human doctor. Algorithmic predictions can be utilized by patients to decide whether to see a doctor for their symptoms. Suppose we have three different algorithms predicting the severity of flu symptoms in patients and would like to decide which one should be deployed in the real world. Each algorithm has a specific level of accuracy—where accuracy specifies the percentage of subjects for whom the algorithm makes a correct prediction. Note that in cases where the deployed algorithm makes an error, a patient will be temporarily impacted negatively. A patient falsely predicted to develop severe flu symptoms may unnecessarily seek medical intervention, while a patient falsely labeled as developing only mild symptoms may have to cope with severe symptoms for a longer period of time (at most two weeks).

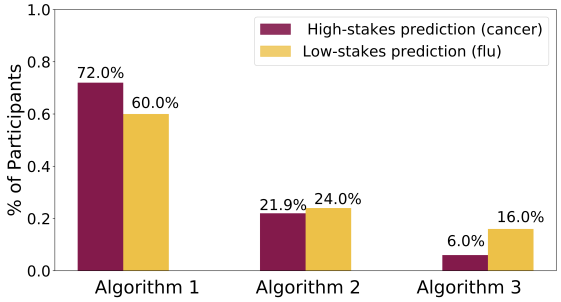

Figure 8: Medical risk prediction scenarios. Participants gave higher weight to accuracy (compared to inequality) when predictions can impact patients' life expectancy.

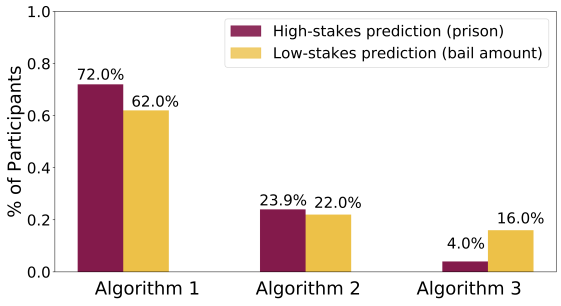

Figure 9: Crime risk prediction scenarios. Participants gave higher weight to accuracy (compared to inequality) when predictions can impact defendants' life trajectory.

Figures 9 and 8 show the number of participants who chose each algorithm. As hypothesized, when the stakes are high, participants gave higher weight to accuracy and lower weight to inequality.

5 DISCUSSION

The main takeaway message from our experiments is that **demographic parity best captures people's perception of fairness.** Our surveys show that **participants consider accuracy more important than equality when stakes are high.** Finally, participants on Amazon Mechanical Turk found the task engaging and informative, while some thought more context about the algorithm and decision subjects would have helped them make a more informed choice. Feedback from the users, such as the examples shown in Table 6, demonstrate that our task encouraged users to think about what factors algorithmic decision-making systems should take into account, as well as reflect about the use of discriminatory algorithms in society.

5.1 Limitations

The primary goal of our work was to find out which existing notion of fairness best captures lay people's perception of fairness in a given context. We took it as a given that at least one of these notions is a good representation of human judgment—at least in the highly stylized setting we presented them with. We acknowledge that real-world scenarios are always much more complex and there are many factors (beyond true and predicted labels) that impact people's judgment of the situation. Existing mathematical notions of fairness are in comparison very simplistic, and they can never reflect all the nuances involved. Our work is by no means a final verdict—it is an initial step toward a better understanding of fairness and justice from the perspective of everyday people.

One barrier to obtaining meaningful answers from participants is engagement. On AMT it is difficult to monitor participants' attention to the task. We were particularly concerned about the possibility of participants choosing their answers without careful consideration of all factors in front of them—and in the worst case, completely at random. We prevented this by

- Restricting participation to turkers with 99% approval rate.
- Limiting the number of tests to at most 20.
- Requiring the participant to provide an explanation for their choice.
- Restricting participants to complete the task at most once.

As with any experiment, we cannot completely rule out the potential impact of framing. We experimented with two different UIs, and our findings were robust to that choice. This, however, does not mean that a different experimental setting could not lead to different conclusions. In particular, in our experiments, participants did not have personal stakes in their choice of algorithm, and they might have responded differently if they could be the subject of decision making through their choice of algorithm.

In all of our tests, we restricted attention to two predictive models of similar accuracy. How do participants' perception of discrimination change if the two models presented to them differ in terms of accuracy? We leave the study of the role of accuracy for future work.

5.2 Future Directions

Our findings have important implications for the Fair ML literature. In particular, they accentuate the need for a more comprehensive understanding of the human attitude toward algorithmic fairness. Algorithmic decisions will ultimately impact human subjects' lives, and it is, therefore, critical to involve them in the process of choosing the right notion of fairness. Our work takes an initial step

Table 6: Examples of participant feedback demonstrating positive educational effect of the task.

Instance	Feedback
1	*"It was more difficult than I first thought to choose an algorithm. The question must come up: is it discriminatory to predict correctly? This is something to mull over. Thank you!"*
2	*"I would have liked to know age and education to see if those were also factors though."*
3	*"This was really fun, thank you! I've never realized that algorithms were used for this sort of data, I hope this helps in some way to improve those systems."*
4	*"The justice system's algorithms think very poorly of white men and black women. That's frightening if true."*
5	*"I thought this was really interesting! I'd be curious what the results say."*

toward this broader agenda. Directions for future work include, but are not limited to, the following:

- providing subjects with more information about subjects of algorithmic decision making and inner workings of each algorithm; quantifying how this additional context affects their assessment of algorithmic fairness.
- providing subjects with information about the non-algorithmic alternatives (i.e., cost, accuracy, and bias of human decisions)
- exploring fairness considerations at different levels; are people more sensitive to individual- or group-level unfairness?
- studying the role of personal stakes; do people assess algorithmic fairness differently if they may be personally affected by algorithmic decisions?
- studying the effect of participant's expertise; does expertise in law and discrimination qualitatively change the participants' responses?

ACKNOWLEDGMENTS

M. Srivastava acknowledges support from the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-1656518 and from the ETH Zürich Student Summer Research Fellowship. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the funding agencies.

REFERENCES

[1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias. *Propublica* (2016).
[2] Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. 2018. The moral machine experiment. *Nature* 563, 7729 (2018), 59.
[3] Anna Barry-Jester, Ben Casselman, and Dana Goldstein. 2015. The New Science of Sentencing. *The Marshall Project* (August 2015).
[4] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. 'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions. In *CHI*. ACM, 377.
[5] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *In Proceedings of the 1st Conference on Fairness, Accountability and Transparency (FAT*)*. 77–91.
[6] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *arXiv preprint arXiv:1703.00056* (2017).
[7] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic decision making and the cost of fairness. In *KDD*. ACM, 797–806.
[8] Rahul C. Deo. 2015. Machine learning in medicine. *Circulation* 132, 20 (2015), 1920–1930.
[9] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the Innovations in Theoretical Computer Science Conference (ITCS)*. ACM, 214–226.
[10] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *KDD*. ACM, 259–268.
[11] Pratik Gajane and Mykola Pechenizkiy. 2017. On formalizing fairness in prediction with machine learning. *arXiv preprint arXiv:1710.03184* (2017).
[12] Daniel Golovin, Andreas Krause, and Debajyoti Ray. 2010. Near-optimal bayesian active learning with noisy observations. In *NIPS*. 766–774.
[13] Nina Grgic-Hlaca, Elissa M. Redmiles, Krishna P Gummadi, and Adrian Weller. 2018. Human perceptions of fairness in algorithmic decision making: A case study of criminal risk prediction. In *WWW*. 903–912.
[14] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. In *NIPS*. 3315–3323.
[15] Robert David Hart. 2017. If you're not a white male, artificial intelligence's use in healthcare could be dangerous. *Quartz* (July 2017).
[16] Hoda Heidari, Michele Loi, Krishna P. Gummadi, and Andreas Krause. 2019. A Moral Framework for Understanding of Fair ML through Economic Models of Equality of Opportunity. In *FAT**.
[17] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudík, and Hanna Wallach. 2018. Improving fairness in machine learning systems: What do industry practitioners need?. In *CHI*.
[18] Matthew Joseph, Michael Kearns, Jamie Morgenstern, and Aaron Roth. 2016. Fairness in learning: Classic and contextual bandits. In *NIPS*. 325–333.
[19] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2017. Inherent trade-offs in the fair determination of risk scores. In *ITCS*.
[20] Min Kyung Lee and Su Baykal. 2017. Algorithmic Mediation in Group Decisions: Fairness Perceptions of Algorithmically Mediated vs. Discussion-Based Social Division.. In *CSCW*. 1035–1048.
[21] Min Kyung Lee, Daniel Kusbit, Anson Kahng, Ji Tae Kim, Xinran Yuan, Allissa Chan, Ritesh Noothigattu, Daniel See, Siheon Lee, and Christos-Alexandros Psomas. 2018. WeBuildAI: Participatory Framework for Fair and Efficient Algorithmic Governance.
[22] Ritesh Noothigattu, Snehalkumar 'Neil' S. Gaikwad, Edmond Awad, Sohan Dsouza, Iyad Rahwan, Pradeep Ravikumar, and Ariel D. Procaccia. 2018. A voting-based system for ethical decision making. In *AAAI*.
[23] Kevin Petrasic, Benjamin Saul, James Greig, and Matthew Bornfreund. 2017. Algorithms and bias: What lenders need to know. *White & Case* (2017).
[24] Debajyoti Ray, Daniel Golovin, Andreas Krause, and Colin Camerer. 2012. Bayesian rapid optimal adaptive design (broad): Method and application distinguishing models of risky choice. *California Institute of Technology working paper* (2012).
[25] Cynthia Rudin. 2013. Predictive Policing Using Machine Learning to Detect Patterns of Crime. *Wired Magazine* (August 2013). Retrieved 4/28/2016.
[26] Nripsuta Saxena, Karen Huang, Evan DeFilippis, Goran Radanovic, David Parkes, and Yang Liu. 2018. How Do Fairness Definitions Fare? Examining Public Attitudes Towards Algorithmic Definitions of Fairness. *arXiv preprint arXiv:1811.03654* (2018).
[27] Till Speicher, Hoda Heidari, Nina Grgic-Hlaca, Krishna P. Gummadi, Adish Singla, Adrian Weller, and Muhammad Bilal Zafar. 2018. A Unified Approach to Quantifying Algorithmic Unfairness: Measuring Individual and Group Unfairness via Inequality Indices. In *KDD*.
[28] Latanya Sweeney. 2013. Discrimination in online ad delivery. *Queue* 11, 3 (2013), 10.
[29] Anne Taylor and Jackson Wright. 2005. Importance of race/ethnicity in clinical trials. *Circulation* 112, 23 (2005), 3654–3660.
[30] Michael Veale, Max Van Kleek, and Reuben Binns. 2018. Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making. In *CHI*. ACM, 440.
[31] Allison Woodruff, Sarah E. Fox, Steven Rousso-Schindler, and Jeffrey Warshaw. 2018. A Qualitative Exploration of Perceptions of Algorithmic Fairness. In *CHI*. ACM, 656.
[32] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. 2017. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *WWW*. 1171–1180.

A REPRODUCIBILITY

A.1 Task Interface

Each of the 20 tests were adaptively selected based on the participant's answers to previous tests, with the first test provided at random. A return code was provided for AMT participants to give back to us after they finished all 20 tests—this allowed us to link AMT participants survey responses (i.e., demographic information and feedback) with the logs of server interactions that we maintained.

A.2 Code

The code for our server, implementation of adaptive test selection, and analysis can all be found on our Git Repository: https://github.com/meghabyte/FairnessPerceptions/. A more detailed description of how to run the code and reproduce the plots in our paper can be found in the repository's ReadMe. A very brief description follows:

Amazon Mechanical Turk data. All raw data from our Amazon Mechanical Turk experiments can be found under the data folder in the repository. These consist of server logs tracking all interactions with the server. We also provide the following processed data files:

- A file combining each participant's data with their likelihoods across hypotheses/explanation after 20 tests;
- A file containing results from our pilot experiments (including the drop-down user interface that we chose to not use in our full experiments).

To respect the participants privacy, we do not release any data consisting of their identifying information, demographics, and feedback.

Analysis. All plots from our paper, as well as the code to regenerate them, can be found in the analysis folder in the repository.

Interface. Our interface can be run locally as well as a test mode where either random test selection (see Figure 3) or random answer generation (see Figure 4) can be simulated. All command line arguments are explained in the repository.

B ADDITIONAL SURVEYS

Sentencing time in prison. *Data-driven decision-making algorithms can be employed to predict the likelihood of future crime by defendants. These algorithmic predictions are utilized by judges to make sentencing decisions for defendants, in particular, how much time the defendant has to spend in prison. Suppose we have three different algorithms predicting the risk of future crime for defendants and would like to decide which one should be deployed in real-world courtrooms. Each algorithm has a specific level of accuracy—where accuracy specifies the percentage of subjects for whom the algorithm makes a correct prediction. See Table 5. Note that in cases where the deployed algorithm makes an error, a defendant's life can be significantly impacted: A defendant falsely predicted to re-offend can unjustly face long prison sentences (minimum 1 year), while a defendant falsely predicted to not reoffend will not spend any time in prison and may commit a serious crime that could have been prevented. From a moral standpoint, which one of the three algorithms do you think is more desirable for deployment in real-world courtrooms?*

Setting the bail amount. *Data-driven decision-making algorithms can be employed to predict the likelihood of a defendant appearing in court for his/her future hearings. These algorithmic predictions can be utilized by judges to set the appropriate bail amount for the defendant. Suppose we have three different algorithms predicting the risk of a defendant not showing up to future court hearings and we would like to decide which one should be deployed in real-world courtrooms. Each algorithm has a specific level of accuracy—where accuracy specifies the percentage of subjects for whom the algorithm makes a correct prediction. Note that in cases where the deployed algorithm makes an error, the defendant may be financially impacted: A defendant falsely predicted to not appear for future hearings is unnecessarily forced to submit a bail amount of $2000. From a moral standpoint, which one of the three algorithms do you think is more desirable for deployment in real-world courtrooms?*