A Human-in-the-loop Framework to Construct Context-dependent Mathematical Formulations of Fairness

Mohammad Yaghini
EPFL
mohammad.yaghini@epfl.ch

Hoda Heidari
Cornell University
hh732@cornell.edu

Andreas Krause
ETH Zürich
krausea@ethz.ch

arXiv:1911.03020v1 [cs.AI] 8 Nov 2019

Abstract

Despite the recent surge of interest in designing and guaranteeing mathematical formulations of fairness, virtually all existing notions of algorithmic fairness fail to be adaptable to the intricacies and nuances of the **decision-making context** at hand. We argue that capturing such factors is an inherently *human* task, as it requires knowledge of the social background in which machine learning tools impact real people's outcomes and a deep understanding of the ramifications of automated decisions for decision subjects and society. In this work, we present a framework to construct *context-dependent* mathematical formulation of fairness utilizing *people's judgment of fairness*. We utilize the theoretical model of Heidari et al. (2019)—which shows that most existing formulations of algorithmic fairness are special cases of economic models of *Equality of Opportunity (EOP)*—and present a practical **human-in-the-loop** approach to pinpoint the fairness notion in the EOP family that best captures people's perception of fairness in the given context. To illustrate our framework, we run human-subject experiments designed to learn the parameters of Heidari et al.'s EOP model (including *circumstance*, *desert*, and *utility*) in a hypothetical recidivism decision-making scenario. Our work takes an initial step toward *democratizing* the formulation of fairness and utilizing human-judgment to tackle a fundamental shortcoming of automated decision-making systems: that the machine on its own is incapable of understanding and processing the human aspects and social context of its decisions.

1 Introduction

Machine Learning (ML) tools are increasingly employed to make consequential decisions for human subjects, in areas such as credit lending (Petrasic et al., 2017), policing (Rudin, 2013), criminal justice (Barry-Jester et al., 2015), and medicine (Deo, 2015). Decisions made by these algorithms can have a long-lasting impact on people's lives and may affect certain individuals or social groups negatively (Sweeney, 2013; Angwin et al., 2016). This realization has recently spawned an active area of research into quantifying and guaranteeing fairness for ML (Dwork et al., 2012; Kleinberg et al., 2017; Hardt et al., 2016).

Despite the recent surge of interest in designing mathematical formulations of fairness, none of the existing notions can adapt to the intricacies and nuances of the *decision-making context* at hand. Designing an ethically acceptable measure of fairness requires knowledge of the social *background* in which data-driven decision-making tools are employed and a deep understanding of the *ramifications* of automated decisions for their subjects and society. There is no one-size-fits-all definition of fairness, and we should not assume that *any* of the existing notions can serve as an ethically acceptable formulation of (un)fairness in a broad range of practical domains. Assuming that an elusive concept like fairness can be reasonably approximated and described in mathematical terms—which itself is a big assumption—we argue that practitioners need to have a framework in place for constructing a *specialized* formalization of fairness *tailored* to the decision-making context.

To further illustrate the issue, consider equality of odds (Hardt et al., 2016)[1] in a hypothetical

[1]Equality of odds requires false positive and false negative rates to be equal for all socially salient groups.

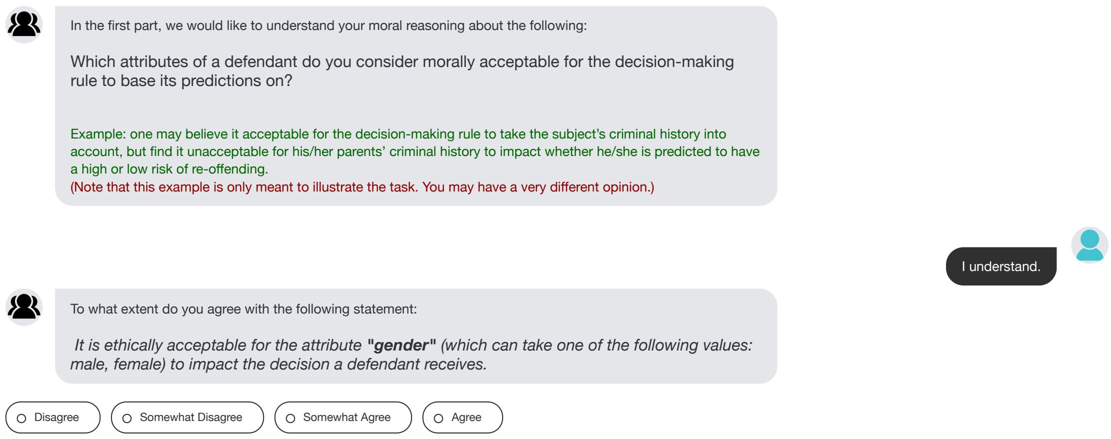

Figure 1: Our conversational user-interface designed to elicit participants' opinion about desert, utility, and circumstance.

recidivism prediction context where predictions directly translate into sentencing decisions. According to (Heidari et al., 2019), one can interpret the equality of odds requirement as follows: all defendants with the same true label deserve the same sentencing outcome regardless of their group membership. In other words, all defendants who would go on to commit another crime if set free should ideally receive the same (high-risk) prediction. Similarly, all defendants who will not commit another crime should receive the same (low-risk) prediction. Now in this context, employing equality of odds as the notion of fairness implies that we agree with the following statement: "The true label is a good proxy for a defendant's *deserved* sentencing outcome, and the predicted label is a good indicator of the *benefit/harm* he/she perceives as the result of their sentencing decision". We argue that we should never make such assumptions unless we have verified them carefully and ensured that they reflect domain-experts and stakeholders' perception of desert and utility in the given context. When it comes to sentencing decisions, for example, we may find out that beyond recidivism, most people are in favor of additionally taking a defendant's age and family situation into account to reach a just sentencing outcome.

In this work, we present a practical framework to *construct* a mathematical formulation of fairness utilizing *people's judgment of fairness* in the application domain at hand. We utilize the theoretical model of Heidari et al. (2019)—which shows that most existing notions of algorithmic fairness are special cases of economic models of *Equality of Opportunity (EOP)*—and present a *human-in-the-loop* approach to pinpoint the most appropriate formulation within the EOP family of fairness notions. To illustrate our approach, we run human-subject experiments on Amazon Mechanical Turk (AMT) designed to learn the parameters of Heidari et al.'s EOP model (that is, circumstance, desert, and utility) for a hypothetical recidivism prediction scenario.

In order to learn the notion that is most compatible with people's perception of fairness, we ask participants to respond to a series of three questionnaires. Each participant is required to answer a total of 55 questions—all concerning our fixed, carefully specified scenario. A typical question in our experiment provides the participant with information about two hypothetical decision subjects (defendants). The participant is then prompted to choose

- which one of the two defendants he/she believes is more *deserving* of receiving a lenient decision?
- which one of them he/she believes would *benefit* more from their algorithmic decision?

Based on the participant's answer to these questions, we can pinpoint the mathematical formulation of fairness that best captures his/her judgment in the decision-making context at hand. (Our user-interface (anonymized for double-blind reviewing) can be found at `https://fair-server.ml/`.) We find that participants are generally in favor of taking factors beyond the defendant's true and predicted labels (e.g., race, gender, age, and criminal history) into account when determining his/her deserved sentencing outcome and perceived benefit/harm. Last but not least, we propose and investigate two methods to aggregate individual participant notions into one socially acceptable definition of fairness. In that sense, our work takes an initial step toward *democratizing* the formulation of fairness for algorithmic decision-making.

We emphasize that attempts to democratize the formulation of fairness should not stop at taking input from *ordinary people* (e.g., participants on AMT) alone, due to numerous ethical considerations (e.g., it is not ethically acceptable for the majority to determine the rights of freedoms of the minority). Practitioners must strive to involve stakeholders, domain experts, community members, ethicists, system designers, and beyond in the process, then carefully deliberate and determine the weight of the input from each party, and decide the circumstances under which their judgments should be overruled (e.g., if they violate the law or ethical norms, or reflect harmful biases and stereotypes). While these tasks are out of the scope of the current paper, they mark critical direction for future work.

In summary, our work draws attention to two important issues: First, fairness is a highly context-dependent ideal and designing an ethically acceptable formulation of fairness requires carefully examining the domain and accounting for the social- and individual-level implications of automated decision-making. Second, Fulfilling these criteria is an inherently human task, and we must utilize human-judgment to tackle a fundamental shortcoming of automated decision-making: that the machine lacks the faculties to understand and process the social and human aspects of its decisions. Our work presents a practical human-in-the-loop toolkit to construct a context-dependent mathematical formulation of fairness and advocates for democratizing the process of formulating fairness for consequential decisions.

1.1 Related Work

Numerous mathematical definitions of fairness have been recently proposed and studied; examples include demographic parity (Dwork et al., 2012), disparate impact (Zafar et al., 2017), equality of odds (Hardt et al., 2016; Feldman et al., 2015), and calibration (Kleinberg et al., 2017). While each of these notions may seem appealing on their own, they are incompatible with one another and cannot hold simultaneously (see (Kleinberg et al., 2017; Chouldechova, 2017)). Furthermore, at least in the context of criminal risk assessment, it is far from settled which one of these notions is a more appropriate measure of algorithmic fairness (see (Angwin et al., 2016) and (Dieterich et al., 2016)).

Several recent papers empirically investigate algorithmic ethics and fairness utilizing human-subject experiments and questionnaires. MIT's moral machine (Awad et al., 2018) provides a crowd-sourcing platform for aggregating human opinions on how self-driving cars should make moral decisions by eliciting their preferences through a series of pairwise preference ordering questions. For the same setting, Noothigattu et al. (2018) propose learning a random utility model of individual preferences, then efficiently aggregating those individual preferences through a social choice function. (Lee et al., 2018) proposes a similar approach for ethical decision-making in the context of food distribution. Similar to the previously-mentioned papers, we obtain input from human-participants by asking them pairwise questions (that is, to compare two alternatives from a moral standpoint). While Noothigattu et al. and Lee et al. focus on modeling human preferences and *aggregating* them utilizing tools from *social choice theory*, our primary goal is to cast human perceptions of fairness as an instance of EOP.

Grgic-Hlaca et al. (2018) study why people perceive the use of certain features as unfair in making recidivism predictions about defendants. Binns et al. (2018) study people's perceptions of justice in algorithmic decision-making under different *explanation styles*. Dressel and Farid (2018) show that the widely used commercial risk assessment software COMPAS is no more accurate or fair than *predictions made by people* (AMT users) with little or no criminal justice expertise. Green and Chen (2019) study the impact of risk assessment tools on the actual decision-making process (e.g., how do judges interpret and use recidivism risk score to make a sentencing decision for a defendant.).

Veale et al. (2018) interview *public sector machine learning practitioners* regarding the challenges of incorporating public values into their work. Holstein et al. (2018) 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 *quantifying* lay people's perception of fairness, Holstein et al. and Veale et al. study ML practitioners' views toward fairness and provide *qualitative* suggestions and best-practices.

Several recent papers in human-computer interaction study users' expectations and perceptions of algorithmic fairness. Lee and Baykal (2017) 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. (2018) 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 attempted to *construct context-dependent* definitions of fairness utilizing human perception of justice. However, two notable articles recently investigate people's attitude towards *existing* mathematical formulations of fairness. Saxena et al. (2018) investigate ordinary people's attitude toward three notions of *individual fairness* in the context of *loan decisions*. Srivastava et al. (2019) investigate which of the *existing* mathematical formulations of *group fairness* best captures people's perception of fairness in two different decision-making scenarios. We emphasize that unlike our work, they do not attempt to *construct* a *context-dependent* measure of fairness utilizing people's feedback.

2 Framework

In this section, we formally define our setting and cast the problem of formulating algorithmic unfairness as that of estimating the parameters of an EOP condition (Heidari et al., 2019). Throughout, we consider the following *simplified* automated decision-making setting: a predictive model (e.g., a hypothetical recidivism prediction model) is trained using historical data records (e.g., the attributes and outcomes of past defendants) through the standard supervised learning pipeline. The predictions made by this model then directly translate into important decisions for never-before-seen subjects (e.g., a future defendant is sent to jail if the model predicts that he/she would re-offend in the future).

More precisely, let $T = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ denote a training data set consisting of n instances, where $\mathbf{x}_i \in \mathcal{X}$ specifies the feature vector for individual i and $y_i \in \mathcal{Y}$, the true label for him/her. Unless otherwise specified, we focus on the binary classification task (i.e., we assume $\mathcal{Y} = \{0, 1\}$) and assume without of loss of generality that $\mathcal{X} = \mathbb{R}^k$. Throughout, we assume that decision subjects consider a positive prediction more desirable than a negative prediction (e.g., defendants prefer bail/parole to jail and a prediction of 1 would correspond to bail/parole).

Individuals in T are assumed to be sampled i.i.d. from a distribution/population F. A learning algorithm uses the training data T to fit a predictive *model* (or hypothesis) $h : \mathcal{X} \rightarrow \mathcal{Y}$ that predicts the label for out-of-sample instances. More precisely, let $\mathcal{H}$ be the hypothesis class consisting of all the models the learning algorithm can choose from. The learning algorithm receives T as the input and selects a model $h \in \mathcal{H}$ that minimizes some notion of empirical loss, $\mathcal{L}(T, h)$, on T. For the ease of notation, when the predictive model in reference is clear from the context, we denote the predicted label for an individual with feature vector $\mathbf{x}$ by $\hat{y}$ where $\hat{y} = h(\mathbf{x})$.

2.1 Fairness as Equality of Opportunity

Heidari et al. (2019) have recently established that existing formulations of (group) fairness, such as equality of odds or statistical parity, can be cast as special cases of economic models of EOP. The core idea at the heart of EOP is to distinguish between factors that can morally justify inequality in outcomes among individuals, and factors that are morally irrelevant and ideally should *not* impact outcomes. An equal opportunity policy is one through which an individual's outcome only depends on the former—and not the latter—types of factors.

Economists have translated EOP into precise mathematical terms (see for instance (Lefranc et al., 2009)). At a high-level, these models break down an individual's attribute into two categories: *circumstance* and *desert*, where circumstance captures all factors that should not affect the individual's *utility*, and desert encapsulates factors that can justify unequal outcomes among individuals. More precisely, let c denote the *circumstance*, capturing factors that are not considered legitimate sources of inequality among individuals (e.g., race in recidivism prediction). Let d denote a *scalar* summarizing factors that are viewed as legitimate sources of inequality (e.g., prior criminal convictions). Let ϕ denote the policy that governs the distribution of *utility*, u, among individuals in society. An individual's utility is a consequence of his/her desert and circumstance and the implemented policy. Formally, let $F^{\phi}(.|c,d)$ specify the cumulative distribution of utility under policy ϕ among people with desert level d and circumstance c. EOP requires that for individuals with similar desert d, the distribution of utility should be the same—regardless of their circumstances. More precisely,

Definition 1 (Equality of Opportunity (EOP)) *A policy ϕ satisfies EOP if for all circumstances c, c' and all desert levels d,*

$$F^{\phi}(.|c,d) = F^{\phi}(.|c',d).$$

It can be shown that existing formulations of fairness for binary classification can be cast as special cases of the above formulation of EOP (Heidari et al., 2019). For instance, if we assume an individual's true label captures their desert, their sensitive group membership specifies their circumstance, and their predicted label is their utility, then we can view Equality of Odds as a special case of the above definition of EOP.

In this work, we restrict our attention to the family of fairness models that can be cast as instances of EOP, and set out to estimate the parameters c, d, and u that best captures the human perception and judgment of fairness in a particular decision-making context.

2.2 Parameter Estimation for EOP

To formulate fairness as EOP, we first need to specify which attributes belong to a subject's circumstance (c), what constitutes his/her deserved outcome (d), and how the decision he/she receives impacts his/her utility (u).

Identifying circumstance c From each participant p, we ask a series of k questions inquiring whether p believes feature $j \in \{1, \cdots, k\}$ can justify inequality in decisions. Let $\mathbf{z}_p$ denote the subset of features that p deems morally irrelevant. We treat each possible value that features in $\mathbf{z}_p$ can take as a distinct circumstance c.

Estimating desert d Consider a participant p and a decision subject with characteristics $\mathbf{x}, y$. Let d_p specify the subject's *overall desert* according to p. We assume d_p is not directly observable, but there exists a function $\delta_p : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}^+$, such that

$$d = \delta_p(\mathbf{x}, y).$$

That is, δ_p maps the information the participant observes about the decision subject (i.e., $\mathbf{x}, y$) to his/her desert. For simplicity, we assume δ_p is a linear function of $\mathbf{x}, y$, that is, there exists a coefficient vector $\boldsymbol{\delta}_p$ such that $d = \boldsymbol{\delta}_p.[\mathbf{x}, y]$ (where $[\mathbf{x}, y]$ is the vector concatenation of feature vector $\mathbf{x}$ and true label y).

To estimate the $\boldsymbol{\delta}_p$, we ask p a series of Q pairwise comparison questions. In each question $q = 1, \cdots, Q$, we present p with the information about two hypothetical decision subjects, $i_1^q = [\mathbf{x}_1^q, y_1^q]$ and $i_2^q = [\mathbf{x}_2^q, y_2^q]$, and ask her which one of them she considers to be more deserving of receiving the positive prediction.[2] We assume p would pick i_1^q with probability

$$\Phi(\boldsymbol{\delta}_p.[\mathbf{x}_1^q - \mathbf{x}_2^q, y_1^q - y_2^q]) \tag{1}$$

where $\boldsymbol{\delta}_p$ is the unknown parameter we wish to estimate and Φ represents the cumulative distribution function of the standard normal distribution. This modeling choice closely matches prior work (Noothigattu et al., 2018).

With this model in place, we can readily find the Maximum Likelihood Estimator (MLE) of $\boldsymbol{\delta}_p$ using p's answers to the pairwise questions presented to her. More precisely, let a^q be p's answer to the q'th pairwise question, in which she has expressed confidence level $c \in \{1, 2\}$. a^q is $|c|$ if p

[2]Note that different participants may receive different sets of questions, hence to be precise, all the notation defined in this part must have a p superscript. For the sake of brevity, when the participant in reference is clear from the context, we drop the superscript p.

chooses subject 1 in response to question q, and it is $-|c|$ otherwise. We find the maximum likelihood estimator or $\boldsymbol{\delta}_p$ by solving the following optimization:

$$\begin{aligned} \arg\min_{\boldsymbol{\delta}} \quad & -\sum_{q=1}^{Q} \log \Phi \left(a^q\, \boldsymbol{\delta}.[\mathbf{x}_1^q - \mathbf{x}_2^q, y_1^q - y_2^q]\right) \\ \text{s. t.} \quad & \|\boldsymbol{\delta}\|_2 \leq 1 \end{aligned} \tag{2}$$

Estimating individual utility u Let u_p denote the *utility* a decision subject $[\mathbf{x}, y]$ earns as a result of being subject to decision rule h according to participant p. For simplicity, we assume u_p is a linear function of $\mathbf{x}, y, \hat{y}$. That is, there exists $\boldsymbol{\upsilon}_p$, such that $u_p = \boldsymbol{\upsilon}_p.[\mathbf{x}, y, \hat{y}]$.

To estimate the $\boldsymbol{\upsilon}_p$, we ask p a series of Q pairwise comparison questions. In each question $q = 1, \cdots, Q$, we present p with the information about two hypothetical decision subjects, $i_1^q = [\mathbf{x}_1^q, y_1^q, \hat{y}_1^q]$ and $i_2^q = [\mathbf{x}_2^q, y_2^q, \hat{y}_2^q]$, and ask her which one of them she believes would benefit more from his/her decision. We assume p would pick i_1^q with probability

$$\Phi(\boldsymbol{\upsilon}_p.[\mathbf{x}_1^q - \mathbf{x}_2^q, y_1^q - y_2^q, \hat{y}_1^q - \hat{y}_2^q]) \tag{3}$$

where $\boldsymbol{\upsilon}_p$ is the underlying parameters we wish to estimate. With this model in place, we can compute the MLE of $\boldsymbol{\delta}_p$ using p's answers to the pairwise questions presented to her.

2.3 Methods of Aggregation

We run our experiments on $N = 99$ participants. For each individual participant p, we estimate the circumstance, desert, and utility functions that best capture his/her judgments. We will denote these as $\mathbf{z}_p, \boldsymbol{\delta}_p$, and $\boldsymbol{\upsilon}_p$, respectively. To aggregate these functions into one representing society as s whole, we consider two natural aggregation methods.[3]

The first method is inspired by Borda Count [4] and it works as follows:

- A feature is part of a decision subject's circumstance if more than half of the participants (or voters) deem it to be so.
- $\boldsymbol{\delta}^*(\mathbf{x}, y) = \frac{1}{N}\sum_{p=1}^{N} \boldsymbol{\delta}_p.[\mathbf{x}, y] = \left(\frac{1}{N}\sum_{p=1}^{N} \boldsymbol{\delta}_p\right).[\mathbf{x}, y]$.
- $\boldsymbol{\upsilon}^*(\mathbf{x}, y, \hat{y}) = \frac{1}{N}\sum_{p=1}^{N} \boldsymbol{\upsilon}_p.[\mathbf{x}, y, \hat{y}] = \left(\frac{1}{N}\sum_{p=1}^{N} \boldsymbol{\upsilon}_p\right).[\mathbf{x}, y, \hat{y}]$.

The second method is inspired by the hierarchical Bayesian modeling approach. We assume there exists a $\boldsymbol{\theta}$ that captures society's perception of desert. A participant p in the society has his own $\boldsymbol{\theta}_p$ which is a noisy version of $\boldsymbol{\theta}$. To estimate $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_p$ we solve the following convex optimization problem:

$$\begin{aligned} \arg\min_{\boldsymbol{\theta}_p, \boldsymbol{\theta}} \quad & -\sum_{p}\sum_{q} \log \Phi \left(a^q\, \boldsymbol{\theta}_p.[\mathbf{x}_1^{p,q} - \mathbf{x}_2^{p,q}, y_1^{p,q} - y_2^{p,q}]\right) \\ \text{s. t.} \quad & \|\boldsymbol{\theta}_p - \boldsymbol{\theta}\|_2 \leq \lambda \\ & \|\boldsymbol{\theta}_p\|_2 \leq 1, \ \|\boldsymbol{\theta}\|_2 \leq 1. \end{aligned} \tag{4}$$

3 Study Design

We conduct a series of questionnaires on AMT to construct a notion of fairness that best captures people's perception of fairness in a hypothetical recidivism prediction context. Note that we illustrate our approach through AMT because of its time- and cost-effectiveness. In real-world applications, obtaining feedback from lay-people alone is by no means sufficient. As noted earlier, practitioners

[3]Many other aggregation methods are conceivable. For brevity, we will only study these two.
[4]See `https://en.wikipedia.org/wiki/Borda_count` for an overview.

must involve stakeholders, domain experts, community members, ethicists, system designers, and beyond in the process, then carefully deliberate and determine the weight of the input from each party.

3.1 Scenarios and Contexts

We first introduce the goals and scope of our study to participants as follows:[5] If the participant needs further information to understand the task, he/she can access a longer version, which includes several examples illustrating the task. See Appendix C.

Background and Task Description Data-driven decision-making algorithms are increasingly employed to automate the process of making important decisions for humans, in areas such as credit lending, medicine, criminal justice, and beyond. **Organizations** in charge of decision-making can utilize massive datasets of historical records to learn a **decision-making** rule capable of making accurate **predictions** about never-before-seen individuals. Such predictions often serve as the basis for consequential **decisions** for these individuals. (From this point on we will refer to **individuals** subject to decisions as **decision subjects**.)
In recent years, several studies have shown that **automated/algorithmic decisions** made in the above fashion may disparately impact certain groups and individuals. For instance, in the context of credit lending, the decision-making rule may systematically disadvantage loan applicants belonging to a certain racial group and reject their loan applications more frequently. These observations have raised many questions and concerns about the fairness of automated decisions.
The goal of our study is to understand your moral reasoning and perception about what it means for automated decisions to be fair—considering the specifics of the decision-making context. We would like to know your ethical judgment through your answers to the following questions:

- Which **attributes** of a decision subject do you consider **morally acceptable** for the decision-making rule to base its decisions on?
- Comparing the attributes of two decision subjects, which one of them do you believe is more **deserving** of receiving a **better (more desirable) decision**?
- How do you believe automated decisions will **impact** these subjects? We would like you to imagine how algorithmic decisions may contribute to the overall **happiness, satisfaction, and well-being** of a decision subject.

Throughout the experiment, we focus on a recidivism prediction context, described below:

Decision-Making Context In court-rooms across the United States, data-driven decision-making algorithms are 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 bail amount, or time to be spent in jail). 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 on bail—compared to a first-time offender, or 2) defendants belonging to certain groups (e.g., residents of neighborhoods with high crime rate) are more likely to reoffend if set free. These automated predictions may directly translate into sentencing decisions. For instance, a defendant who is predicted to have a high risk of reoffending may be sentenced to jail, whereas a defendant who is predicted to have a low risk of reoffending may be set free on bail.

3.2 User Interfaces

Through our *conversational interface*[6] the participant then responds to three questionnaires. The first questionnaire contains five questions, and the second and third parts, each consists of 25 questions.

[5] Throughout, text boxes display the content shown to our study participants through the user interface, *verbatim*.
[6] Conversational UIs are known to increase the user's attention and feel more natural to work with. See (Rice, 2019).

Once the participant completes the main three questionnaires, they can optionally provide us with their demographic information. (Participants had the option to not respond to the questions in this part, and their compensation was not be impacted by how they responded to this exit questionnaire).

Part 1: Identifying z To determine which attributes of a decision subject belong to the category of morally irrelevant, the participant first responds to a simple questionnaire consisting of 5 questions of the following form:

To what extent do you agree with the following statement: It is ethically acceptable for the attribute [...] (which can take one of the following values: [...]) to impact the decision a defendant receives.

For an example of this type of question, see Figure 2. The participant can choose their response from a 4-point Likert scale (“Disagree”, “Somewhat Disagree”, “Somewhat Agree”, and “Agree”). The participant was encouraged (but not required) to provide a justification for their choice in a free-form text.

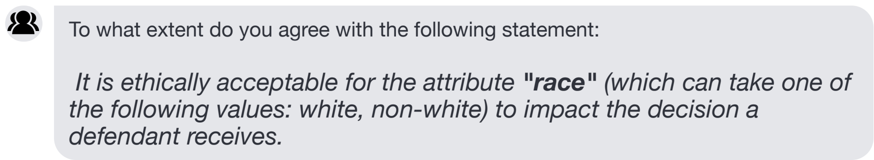

Figure 2: A typical question in part 1 of our experiments.

Before starting the questionnaire in this part, we show the following introductory text to the participant:

In the first part, we would like to understand your moral reasoning about the following: Which attributes of a defendant do you consider morally acceptable for the decision-making rule to base its predictions on? **Example:** one may believe it acceptable for the decision-making rule to take the subject’s criminal history into account, but find it unacceptable for his/her parents’ criminal history to impact whether he/she is predicted to have a high or low risk of reoffending. (Note that this example is only meant to illustrate the task. You may have a very different opinion.)

Part 2: Estimating δ To estimate the desert function (i.e.,$\boldsymbol{\delta}$), we ask the participant to respond to a series of pairwise comparison questions of the following form:

From an ethical standpoint, between the following two decision subjects, who do you believe deserves a more lenient decision?

For an example of this type of question, see Figure 3. The participant can choose either “Clearly subject 1”, “Possibly subject 1”, “Possibly subject 2”, and “Clearly subject 2”. The participant was encouraged (but not required) to provide a justification for their choice in a free-form text.

Before starting the questionnaire in this part, we show the following introductory text to the participant:

From an ethical standpoint, between the following two decision subjects, **who do you believe deserves a more lenient decision**?

ATTRIBUTE	SUBJECT #1	SUBJECT #2
Age Category	Younger than 25	Older than 25
Race	Non-white	Non-white
Gender	Male	Male
Charge Degree	Felony	Felony
Prior Counts	4	9
Actual Outcome	Will reoffend	Will reoffend

Note: The decision subject differences are marked in blue. If you are unsure about the meaning of any attribute, hold the cursor on it to see a definition.

Figure 3: A typical question in part 2 of our experiments.

> In the second part, we would like to understand your moral reasoning about the following: Comparing the attributes of two defendants, which one of them do you believe is more deserving of receiving a more lenient decision?
> **Example:** Consider two defendants with similar attributes, except for their employment status—one unemployed, the other a local government employee. One may consider the employed subject more deserving of the "low risk to reoffend" prediction.
> (Note that this example is only meant to illustrate the task. You may have a very different opinion.)

Part 3: Estimating $\boldsymbol{v}$ To estimate the utility function (i.e., $\boldsymbol{v}$), we ask the participant to respond to a series of pairwise comparison questions of the following form:

> From an ethical standpoint, between the two following decision subjects, who do you think will benefit more from their algorithmic decision?

For an example of this type of question, see Figure 4. The participant can choose either "Clearly subject 1", "Possibly subject 1", "Possibly subject 2", and "Clearly subject 2". The participant was encouraged (but not required) to provide a justification for their choice in a free-form text.

From an ethical standpoint, between the two following decision subjects, **who do you think will benefit more from their algorithmic decision?**

ATTRIBUTE	SUBJECT #1	SUBJECT #2
Age Category	Older than 25	Older than 25
Race	Non-white	White
Gender	Male	Male
Charge Degree	Misdemeanor	Felony
Prior Counts	4	4
Algorithmic Decision	Low risk to reoffend	Low risk to reoffend
Actual Outcome	Will not reoffend	Will not reoffend

Note: The decision subject differences are marked in blue. If you are unsure about the meaning of any attribute, hold the cursor on it to see a definition.

Figure 4: A typical question in part 3 of our experiments.

Before starting the questionnaire in this part, we show the following introductory text to the participant:

In the third part, we would like to understand your moral reasoning about the following: Given the attributes of two defendants, which one of them do you believe would benefit more from their respective algorithmic decision? In responding to this question, imagine yourself in the circumstances of these two defendants and think about how the sentencing decision they receive may affect their lives.
Example: Consider two defendants with similar attributes, except for their number of dependants (one with two children and another with no dependents.). One may believe that a "low risk to reoffend" prediction would contribute more to the overall satisfaction, happiness, and well-being of the subject who has kids.
(Note that this example is only meant to illustrate the task. You may have a very different opinion.)

3.3 Experimental Design

We used a summarized and pre-processed version of the COMPAS data set (Angwin et al., 2016) to generate the questions in our experiment. We restricted attention to the following attributes for each defendant: **gender** (1 if the defendant is male, 0 otherwise), **age** (1 if the defendant is less than 25, 0 otherwise), **race** (1 if the defendant is not Caucasian, 0 otherwise), **charge degree** (1 if felony, 0 otherwise), **prior counts** (a positive integer), **true label** (1 if the defendant reoffends within 2 years, 0 otherwise), and **predicted label** (1 if the COMPAS tool gives a score of 5 or higher[7] to the defendant, 0 otherwise).

In part 1, we asked one question for each of the five features in the data set. In part 2, to generate a question q, we picked the first defendant (i_1^q) i.i.d. from the data set, then chose the second defendant (i_2^q) such that it does not differ with the first one in more than two attributes. The reason for this restriction was two-fold: (1) As shown in Figure 5 it allows us to recover the underlying weight vector confidently with a smaller number of questions; (2) it reduces the cognitive burden of comparing the two defendants with one another for the participants[8]. Questions in part 3 were generated similarly.

Part 2 and 3 each consisted of 25 questions. We chose the number of questions by simulating our model of participants responses (Equation 1) and finding the number of questions necessary to learn their weight vectors with a satisfactory accuracy (i.e., 90%). See figure 5.

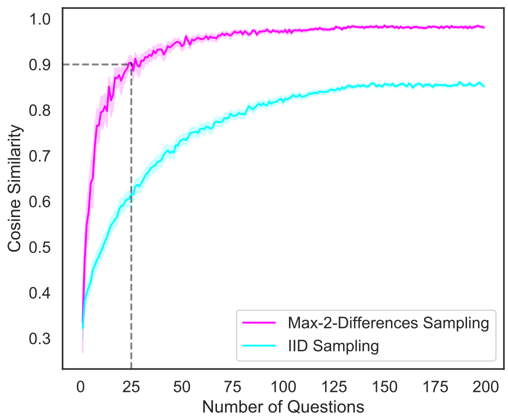

Figure 5: Cosine similarity between the true weight vectors (according to which responses are simulated) and the estimated ones (MLE computed over simulated responses).

[7]Note that the COMPAS tool bases its risk scores on many features other than the ones considered here. Our focus in this work is not on the COMPAS tool and its inner workings. We simply use the COMPAS risk scores (provided by Angwin et al. (2016)) to generate binary predictions for defendants in our hypothetical scenario.

[8]Comparing two objects that differ across multiple dimensions is a type of *multitasking*—which notoriously reduces human attention and leads to erroneous responses. See `https://en.wikipedia.org/wiki/Human_multitasking` and the references therein.

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). Based on the feedback we received from pilot participants, we made several (minor) changes to the text to improve readability and added the option to edit responses. We also decided to randomize the order in which questionnaires in part 2 and 3 are shown to the participants, and included one attention-check question in both.

4 Findings

Through our experiments on AMT, we gathered a data set consisting of 99 participants' responses to our questionnaires. In the end, 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 systematic variations in perceptions of fairness across different demographic groups.[9] Table 1 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 to a particular segment of the population who have Internet access and are willing to complete crowd-sourced tasks online). In particular, for our experiments, participants were younger and more liberal compared to the average U.S. population.

Table 1: Demographics of our AMT participants compared to the 2016 U.S. census data.

Demographic Attribute	AMT	Census
Male	61%	49%
Female	39%	51%
Caucasian	73%	61%
African-American	9%	13%
Asian	10%	6%
Hispanic	4%	18%
Liberal	65%	33%
Conservative	24%	29%
High school	34%	40%
College degree	58%	48%
Graduate degree	8%	11%
18–25	12%	10%
25–40	65%	20%
40–60	17%	26%

4.1 Quantitative Results

Part 1 results: Identifying morally irrelevant features Figure 6 shows the distribution of responses to questions in part 1 of our experiment. Responses represent participants' extent of agreement with the statement "It is ethically acceptable for the attribute [...] to impact the decision a defendant receives." The majority of our participants strongly believe the number of prior counts and charge degree can morally justify inequality in sentencing decisions. The majority also strongly believes race is morally irrelevant. For gender and age, opinions are more diverse. Overall, participants consider gender as part of a defendant's circumstance, and age as a morally relevant feature, however, their degree of confidence is generally lower compared to the other three features.

Part 2 results: Estimating the desert function (δ) Figure 13 (left) shows the distribution of weights given to the five features of a defendant (race, gender, age, charge degree, and prior counts)

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

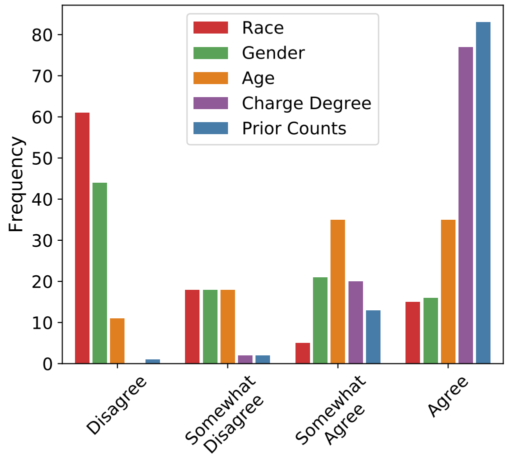

Figure 6: The distribution of responses to questions in part 1 (extent of agreement with the statement "It is ethically acceptable for the attribute [...] to impact the decision a defendant receives.") According to our participants, the number of prior counts, charge degree, and age can morally justify inequality in sentencing decisions, while race and gender are considered morally irrelevant.

when estimating desert ($\boldsymbol{\delta}$). (For ease of interpretation, a Gaussian probability density function with the same mean and variance as the data is plotted on the opposite axis with bolder typeface. We have used 20 bins for these illustrations.) According to our participants, younger defendants generally deserve a more lenient sentencing decision. The same holds if a defendant is female or non-white. Participants consider a defendant less deserving of a lenient outcome if they are charged with a felony, or they have a large number of prior convictions. Perhaps surprising, participants, on average, gave a larger weight (-0.25) to charge degree—an *alleged* offense—as opposed to prior *conviction* counts (-0.46).

Figure 7 (left) illustrates the distribution of weights allocated to the true outcome y when estimating $\boldsymbol{\delta}$. As one may expect, the true outcome has a significant weight, but somewhat surprisingly, on average participants gave similar weights to charge degree (-0.46) and the true label (-.41).

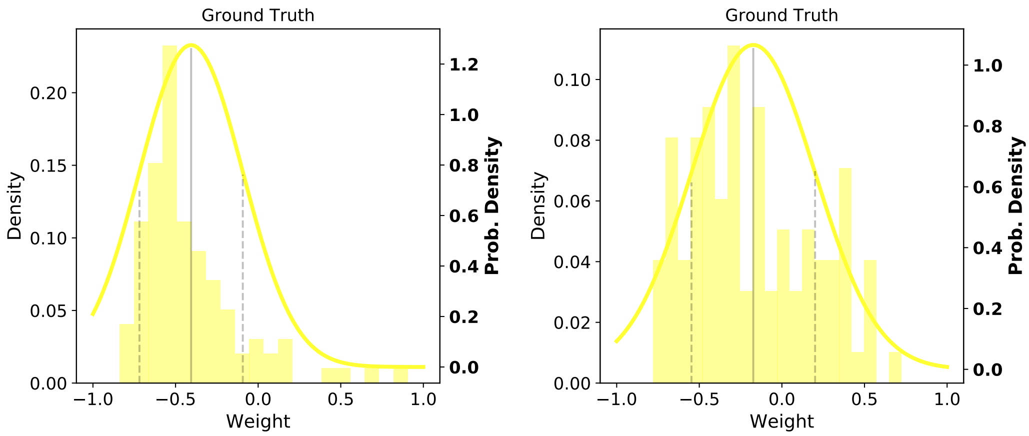

Figure 7: (Left) The distribution of δ for the true label (y) when $\boldsymbol{\delta}$ is estimated for each of our 99 participants, separately. (Right) The distribution of υ for the same setup.

Part 3 results: Estimating the utility function ($\boldsymbol{\upsilon}$) Figure 13 (right) shows the distribution of weights given to the five features (race, gender, age, charge degree, and prior counts) when estimating utility ($\boldsymbol{\upsilon}$). Here, trends are similar to $\boldsymbol{\delta}$ weights, although the spreads are generally larger. Unlike the case of desert, charge degree and prior counts have similar contributions to the overall utility of a defendant.

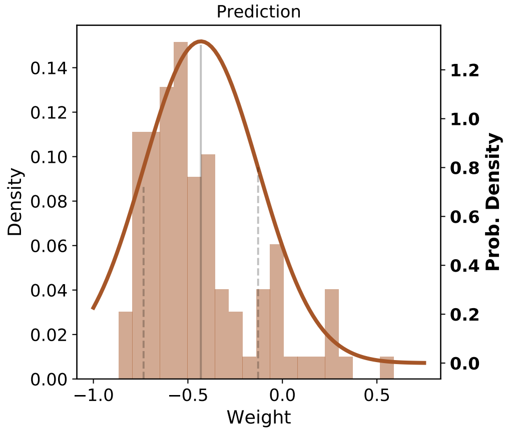

Figure 8: The distribution of υ for the predicted label ($\hat{y}$) when $\boldsymbol{\upsilon}$ is estimated for each of our 99 participants, separately.

Figure 7 in the Appendix (left) illustrates the distribution of weights allocated to the true outcome y when estimating $\boldsymbol{\upsilon}$ and Figure 8 shows the distribution of weight on the predicted outcome, $\hat{y}$. The contribution of $\hat{y}$ is significantly larger than other factors.

Table 2 summarizes the average weights obtained by estimating $\boldsymbol{\delta}$ and $\boldsymbol{\upsilon}$. Note that while the majority of participants consider race and gender as morally irrelevant in part 1, on average, they believe these factors should play a role in determining a defendant's deserved sentencing decision and the harm/benefit they perceive as a result of their sentencing outcomes.

Table 2: The average coefficient of each attribute in our participants' estimated desert and utility functions.

	Race	Gender	Age	Charge	Prior	y	$\hat{y}$
δ	0.11	-0.19	0.13	-0.46	-0.25	-0.41	N/A
υ	0.06	-0.07	0.09	-0.09	-0.08	-0.18	-0.43

Variation across gender, race, age, education, and political views Tables 3, 4 shows the average weights allocated to each attribute of a defendant when $\boldsymbol{\delta}$ and $\boldsymbol{\upsilon}$ are computed for various demographic subgroups of participants. In terms of desert ($\boldsymbol{\delta}$), liberal participants give a relatively large positive weight to race, while conservative participants give it a weight close to 0 (indicating that they believe race does not affect whether a defendant deserves leniency). A similar trend holds for female vs. male, non-white vs. white, and young vs. old participants. Younger participants allocate a large positive weight to age (indicating that they believe younger defendants are generally more deserving of leniency), while our older participants allocate a small negative weight to age.

In terms of harm/benefit to defendants, conservative participants believe female defendants benefit more from leniency, while liberal participants give a relatively small (but still negative) weight to gender. Similar to the case of desert, younger participants allocate a large positive weight to age, while older participants give a small negative weight to age.

Sensitivity to the aggregation method Figure 9 contrasts the weights computed according the two aggregation methods described in Section 2.3. While the weights are generally comparable, we observe that the hierarchical model puts a significantly large weight on prior counts. This is to be expected as this aggregation method zeros in on the points of agreement among participants, and indeed, the majority of participants believe that a defendant with a large prior count is less deserving of leniency.

Goodness-of-fit comparison with an existing notion of fairness Next, we compute the log-likelihood of observing the responses from each participant assuming he/she follows a commonly-

Table 3: Average $\boldsymbol{\delta}$ for participants belonging to various demographic segments. Significant differences are highlighted.

Participant	Race	Gender	Age	Charge	Prior	y
Liberal	**0.16**	−0.22	0.15	−0.46	−0.27	−0.43
Conservative	**0.01**	−0.15	0.10	−0.45	−0.22	−0.34
High-school	0.08	−0.18	0.11	−0.51	−0.31	−0.40
University	0.13	−0.20	0.14	−0.44	−0.22	−0.41
Male	**0.06**	−0.20	**0.17**	−0.42	−0.27	−0.41
Female	**0.19**	−0.18	**0.07**	−0.53	−0.21	−0.40
White	**0.09**	−0.19	0.11	−0.48	−0.28	−0.45
Non-white	**0.19**	−0.23	0.20	−0.41	−0.17	−0.29
Young (¡40)	**0.14**	−0.20	0.18	−0.44	−0.24	−0.41
Old	**0.01**	−0.17	**-0.03**	−0.53	−0.30	−0.39

Table 4: Average $\boldsymbol{\upsilon}$ for participants belonging to various demographic segments. Significant differences are highlighted.

Participant	Race	Gender	Age	Charge	Prior	y	$\hat{y}$
Liberal	0.05	**-0.06**	0.08	−0.10	−0.08	**-0.20**	−0.47
Conservative	0.07	**-0.17**	0.10	−0.09	−0.08	**-0.07**	−0.29
High-school	**0.13**	−0.09	0.13	−0.10	−0.08	−0.23	−0.32
University	**0.02**	−0.06	0.06	−0.09	−0.08	−0.15	−0.48
Male	0.06	−0.04	0.07	−0.06	−0.08	**-0.13**	−0.46
Female	0.05	−0.12	0.11	−0.14	−0.09	**-0.26**	−0.38
White	0.06	−0.06	0.09	**-0.05**	−0.07	−0.16	−0.42
Non-white	0.05	−0.10	0.07	**-0.20**	−0.11	−0.23	−0.45
Young (¡40)	0.04	−0.09	**0.13**	**-0.12**	−0.08	−0.18	−0.43
Old	0.10	0.00	**-0.06**	**0.00**	−0.09	−0.17	−0.40

studied notion of fairness (namely, Equality of Odds) and compare it with our approach (which finds the desert and utility vectors that maximize the likelihood of observing his/her responses. Recall that Equality of Odds assumes the true label indicates desert (in other words, the $\boldsymbol{\delta}$ corresponding to equality of odds allocated a non-zero weight on y, and 0 to all other attributes.) With this assumption, the log-likelihood of observing our participants' responses is on average -27.17, which is significantly worse than our approach with a log-likelihood of -7.3. For the utility function, these numbers are -26.7 and -9.17 for equality of odds and our approach, respectively.

4.2 Qualitative Results

Overall, most participants thought race and gender should not be a deciding factor in determining a defendant's sentencing outcome, even if they are statistically correlated with certain types of crimes. Opinions about age were mixed. Some thought younger people are less in control of their decision, so they should be treated with leniency. Others thought a punishment at a younger age is more effective at preventing future crimes. (See Appendix A for the quotations supporting our statements in this section.)

The majority of participants thought prior convictions should clearly affect a defendant's sentencing outcome. Although, some noted that in a biased judicial system, certain races might have systematically higher numbers of prior convictions.

Recall that in addition to race, gender, age, charge degree and prior convictions, in part 2 we provided the participants with the defendants' true labels (i.e., whether the defendant will reoffend in the future). While they generally thought a participant who will not reoffend in the future is more deserving of leniency, some noted that such information is not available at the time of decision-making,

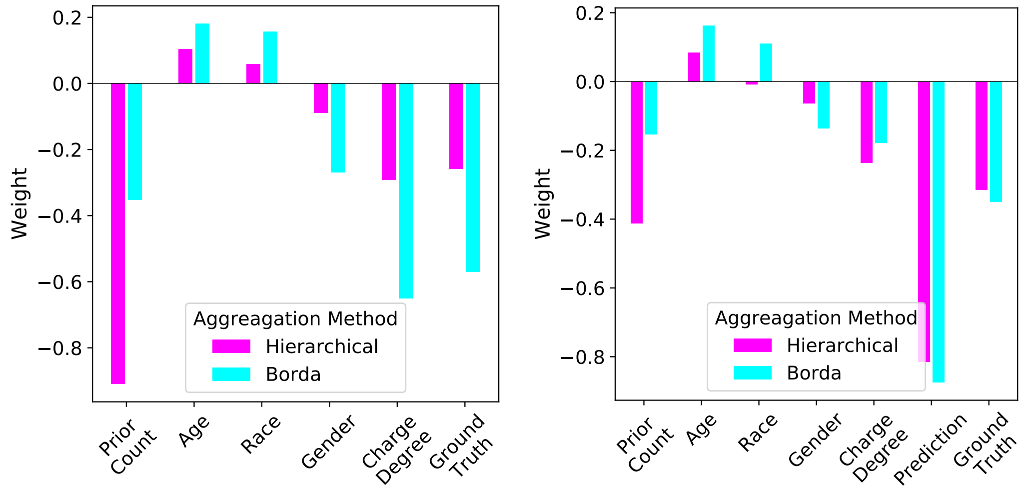

Figure 9: (Left) $\boldsymbol{\delta}$ for the two aggregation methods proposed in Section 2.3. (Right) $\boldsymbol{\upsilon}$ for the two methods.

so neither defendant should get preferential treatment.

In part 3, we provided the participants with the defendants' predicted labels (i.e., whether the algorithm predicts the defendant has a high risk of reoffending in the future). Participants generally agreed that a low-risk prediction is more beneficial for defendants, while a high-risk prediction is particularly harmful if the defendant does not have a long criminal history and is not going to reoffend.

Overall, participants found our task engaging and thought-provoking. For instance, one participant said *"[I] thought it was very interesting and wonder how much an algorithm vs. the human mind really goes into play in possibly life-changing decisions!?!"* Some noticed that "in some cases an algorithm might really be unfair" and hoped that the algorithm takes more than 6-7 attributes into account in making consequential sentencing decisions.

Several participants mentioned that they enjoyed the conversational survey format. Several quotations from our participants follow: (1) *"Great job. one of the best surveys I've been a part of on MTurk."* (2) *"This was such an interesting layout, and the pay was outstanding."* (3) *"This is a very unique survey [...], I'd like to see more in this style."* They also appreciated the option to provide us with feedback about the task.

In summary, our results showcase the thought-provoking and nuanced reasoning of our participants about desert and utility: Our participants took race and gender into account when determining a defendant's desert and utility. For instance, if a defendant is female, participants generally consider her more deserving of a lenient sentencing outcome (compared to an identical male defendant). As another example, participants generally believe the utility of staying out of jail is higher for a young defendant (compared to an older but otherwise identical defendant). These beliefs are aligned with the participants' perception of the social and economic consequences of sentencing decisions for female and young defendants, respectively (as reflected in their justifications). The insights offered by our experiment can have important implications for the formulation of algorithmic fairness in practice.

5 Discussion

Our framework allows practitioners to estimate circumstance, desert, and utility using human judgment as input. These estimates specify a context-dependent formulation of fairness. The resulting notion of fairness can then be utilized like any other mathematical formulation of fairness (e.g., as a constraint in the empirical risk minimization program).

Our small-scale illustration shows the importance of taking the decision-making context into account when designing a mathematical formulation of fairness. In the context of recidivism prediction, our participants believed that beyond the actual outcome (whether the defendant actually goes on to re-offend) and recidivism prediction (whether the defendant is predicted to re-offend; that is, their

COMPAS risk score is higher than 5), other factors such as age, seriousness of the crime, and prior criminal history must be taken into account for algorithmic decisions to be fair.

5.1 Limitations

Focus on the EOP family of notions The primary goal of our work was to construct a context-dependent mathematical formulation of fairness by obtaining people's responses to pairwise questions and finding the (EOP-based) fairness notion that best captures those responses. We took it as a given that the EOP family of fairness notions can represent the human judgments of fairness reasonably well—at least in the hypothetical scenario we presented to our participants. We acknowledge that real-world scenarios are always much more complex, and there are often quite a few factors that impact people's judgment of the situation. We acknowledge that no computationally tractable mathematical formulation of fairness can ever reflect all the nuances involved in achieving fairness, but that does not justify picking an arbitrary formulation without a careful ethical analysis.

Engagement on AMT One barrier to obtaining meaningful answers from participants is to maintain participants' engagement and attention to the task. To prevent the possibility of participants choosing their answers without considering the information provided to them—and in the worst case, completely at random, we took the following steps:

- Restricting participation to Turkers with 99% approval rate.
- Limiting the number of questions to the necessary minimum.
- Incentivizing the participant to justify their answers.
- Adding randomly-placed attention-check questions to our questionnaires.
- Restricting participants to complete the task at most once.

Framing Bias As with any experiment, we cannot completely rule out the potential impact of framing. We attempted to limit this by ensuring a neutral language throughout and reminding the participant they may have a different opinion every time we presented them with a subjective example to demonstrate the task.

Linear models We assumed that both desert and utility functions have a simple linear form.These restrictions allowed us to simplify and scale our illustration, but we fully acknowledge that in practice, such choices have ethical implications that need to be thoroughly analyzed. Linear models simplified our experiments because (1) they are easy to train with a small sample size without the risk of overfitting, and (2) are interpretable. As shown in Figure 5, we can learn the coefficients of a linear preference vector (up to a constant) by asking only 25 questions on average from each participant. Second, linear functions are among the most human-interpretable models. Given that one of the central premises of our work was to bring people in the process, we believe interpretability is an essential feature for the adoption and integration of our framework and fairness measures in practice.

Limited number of features We selected the specific set of observable features from the COMPAS data set for the following two reasons. First, participants could easily understand these features. Second, this small subset of attributes would cover a wide range of moral and social relevance. We expected some of them to be considered arbitrary by the majority of our participants (e.g., race), some highly relevant (e.g., prior counts), and some in between (e.g., age). In practical situations, there must be a thorough deliberation on what features should contribute to each EOP-parameter and in what form.

The standard Gaussian CDF for noise Note that when modeling human behavior, we cannot expect the model to predict the behavior of the human subject, deterministically. Instead, the model must be able to accommodate noisy observations. We utilized a standard Gaussian/normal distribution for noise because it is a common choice in modeling individual pair-wise choices and has been previously adopted in the literature. In practice, if domain-knowledge prescribes a specific distribution for noise, that distribution should replace the normal CDF used in our illustration.

Our use of the COMPAS dataset Our use of the COMPAS data set was solely for the purpose of illustrating our framework on a dataset grounded in the real-world. We made several simplifying assumptions about the hypothetical decision-making context and the dataset. For instance, we assumed the dataset is unbiased. Dealing with the biases encoded in this specific dataset—while highly important—falls out of the scope of this work. As another example, we provided a crude definition for features, such as prior counts, to our participants . Importantly, we did *not* distinguish between different types of prior convictions—in part to keep the survey simple and in part because we did not have access to such fine-grained data.

Biased training data Throughout, we focus on a simplified setting in which we assume the true label is unbiased. In real-world settings (e.g., in case of COMPAS), the training data itself may be biased. While our framework would allow participants to partially address such biases (e.g., by adjusting the coefficient of race in the desert function to partially combat the existing biases against certain races in the judicial system), addressing the biases in the data is in itself a substantial challenge that falls out of the scope of this work.

The absence/presence of a neutral response We intentionally did not provide a neutral response option to our participants to block them from taking a neutral moral stance. Instead, we provided them with the option to verbalize their opinion through free-form text boxes after each question. We decided against providing a neutral option in our main experiment for two reasons. First, we worried that it would dissuade the participants from an in-depth analysis of the dilemma put in front of them (i.e., comparing the two defendants in terms of desert or utility). Second, and on the technical side, a neutral response would not provide the model with any additional information, forcing us to increase the number of questions. We repeated our experiments (at a smaller scale) including a neutral response option in the user-interface, and did not observe significant differences in the estimated weight vectors. Our findings can be found in Appendix B.

On the incomparability of defendants Note that throughout our experiments, participants had the option to provide a written free-form response to questions. The number of cases in which they indicated that they could not make the comparison was negligible.

Small sample size While we provide a separate analysis of the data gathered from different demographic groups, we emphasize that because of the small sample size, the obtained results do not necessarily generalize to the entire population and are for illustration purposes only.

Hypothetical context Last but not least, we focused on a simplified hypothetical context to illustrate our framework. We hope that further experiments in other real-world decision-making contexts will further establish the power of our framework in constructing a context-dependent formulation of fairness.

5.2 Future Directions

Directions for future work include, but are not limited to, the following:

- Providing subjects with more information about decision subjects and inner workings of the decision-making algorithm, and investigating how this additional information affects their assessment of algorithmic fairness.
- Investigating the role of personal stakes: do people assess algorithmic fairness differently if they are personally subject to automated decision-making?
- Studying the effect of participant's expertise (for instance, does knowledge of ethics or law systematically change the participants' responses? What about domain expertise?)

In the near future, we plan to publish a (non-paid) version of our experiment online, where the public can respond to our questionnaires, obtain their results, and compare them with other participants. We believe this will raise awareness around algorithmic decision-making, encourages people to think critically about automated decisions and gain further insight into their mental model of fairness.

References

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias. *Propublica* (2016).

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.

Anna Barry-Jester, Ben Casselman, and Dana Goldstein. 2015. The New Science of Sentencing. *The Marshall Project* (August 2015).

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.

Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *arXiv preprint arXiv:1703.00056* (2017).

Rahul C. Deo. 2015. Machine learning in medicine. *Circulation* 132, 20 (2015), 1920–1930.

William Dieterich, Christina Mendoza, and Tim Brennan. 2016. COMPAS risk scales: Demonstrating accuracy equity and predictive parity. *Northpoint Inc* (2016).

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

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.

Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *KDD*. ACM, 259–268.

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*. ACM, 90–99.

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.

Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. In *NIPS*. 3315–3323.

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**.

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*.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2017. Inherent trade-offs in the fair determination of risk scores. In *ITCS*.

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.

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.

Arnaud Lefranc, Nicolas Pistolesi, and Alain Trannoy. 2009. Equality of opportunity and luck: Definitions and testable conditions, with an application to income in France. *Journal of Public Economics* 93, 11-12 (2009), 1189–1207.

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*.

Kevin Petrasic, Benjamin Saul, James Greig, and Matthew Bornfreund. 2017. Algorithms and bias: What lenders need to know. *White & Case* (2017).

Bill Rice. 2019. Conversational UI is Becoming the Preferred Consumer Interface: 12 Reasons Why. `https://clinc.com/conversational-ui-preferred-consumer-interface/`. *Clinc* (2019).

Cynthia Rudin. 2013. Predictive Policing Using Machine Learning to Detect Patterns of Crime. *Wired Magazine* (August 2013). Retrieved 4/28/2016.

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).

Megha Srivastava, Hoda Heidari, and Andreas Krause. 2019. Mathematical Notions vs. Human Perception of Fairness: A Descriptive Approach to Fairness for Machine Learning. *arXiv preprint arXiv:1902.04783* (2019).

Latanya Sweeney. 2013. Discrimination in online ad delivery. *Queue* 11, 3 (2013), 10.

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.

Allison Woodruff, Sarah E. Fox, Steven Rousso-Schindler, and Jeffrey Warshaw. 2018. A Qualitative Exploration of Perceptions of Algorithmic Fairness. In *CHI*. ACM, 656.

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 Justifications

A.1 Response to Part 1

Race: Overall, most participants thought race should not be a factor in determining a defendant's sentence, even if it is statistically correlated with certain types of crimes. The following quotes from our participants are meant to show the diversity of opinions regarding race.

- *"There is no justification for using race as a factor even if it may have statistical significance."*
- *"That would sort of be like racially profiling which i don't think is fair."*
- *"Using this will only lead to bigger societal problems."*
- *"Race shouldn't be a sole deciding factor, but can be used to make a more accurate decisions for re-offenders of certain crimes."*
- *"Some races have been discriminated against which has led to higher crime, this would just accentuate that problem."*
- *"While majority of people in prison are minorities, I feel it's not fair because it could hurt people who are already disadvantaged."*
- *"I think one's economic position does impact sentencing but not race per se."*
- *"absolutely, certain groups/races have more of the possibility to re-offend."*

Gender: Overall, most participants thought gender should not be a factor in determining a defendant's sentence, even if it is statistically correlated with certain types of crimes. The following quotes from our participants are meant to show the diversity of opinions regarding gender.

- *"Males and females should always expect to receive the same version of justice."*
- *"Gender is something inherent about a person that cannot be changed. while statistically there are likely differences in commission of crimes by men vs. those committed by women, it is unfair to judge someone on that basis."*
- *"Gender is a fluid type of classification, people can even change genders if they wanted to."*
- *"Statistically i think that more men offend or re-offend and if that is the case, I think that is acceptable to include this trait in the decision-making rules."*
- *"[Gender] is part of the criminal, so it certainly should be considered."*

Age: Opinions about age were mixed. Some thought younger people are less in control of their decision, so they should be treated with leniency. Others thought a punishment at a younger age is more effective at preventing future crimes. The following quotes from our participants are meant to show the diversity of opinions regarding gender.

- *"The criteria should be about cognitive ability, not age, as such. young people are believed less responsible for decision-making and the justice system is and should consider that in decisions."*
- *"If they are younger it means their brain isn't fully developed and you should consider that when evaluating their decision making process."*
- *"Minor crimes should have a lesser charge but major crimes should not take age into account. [Leniency toward the youth] carries a risk as well of younger people thinking they can do what they want with less of a punishment when in fact they age at an age where the punishment is more likely to teach them a lesson."*
- *"Older people are probably less likely to reoffend."*

- *"An older suspect would, on average, live less years, so would have less chance to reoffend."*
- *"Good to make an impact at an early age so one does not reoffend."*

Prior Counts: The majority of participants thought prior convictions should clearly affect a defendant's sentencing. Although, some noted that in a biased judicial system, certain races may have systematically higher numbers of prior convictions. The following quotes from our participants are meant to show the diversity of opinions regarding prior counts.

- *"If someone can commit harsh crimes, they probably don't care about others or themselves and will re-offend."*
- *"If one has been convicted before but continues breaking the law, then it is logical to assume this will continue."*
- *"Previous offenses is directly about that specific person's history and not a general data set based on a group (age, sex, race, etc), so I believe it's fair to judge someone based on their own previous history."*
- *"Prior counts are a result of the racially biased criminal justice system. If something could be done to counterbalance inherent racism in the training data, then prior counts could be useful and ethically acceptable."*

Charge Degree: Similar to prior counts, most participants thought charge degree must be an important factor in sentencing decisions.

- *"A serious crime which involves violence especially requires a longer and more severe restriction on the defendant."*
- *"If the racially biased training data could be avoided, then charge degree would be ethically acceptable to use in the system."*

A.2 Response to Part 2

The justifications provided based on the defendant's attributes, such as race, age, and prior counts, generally reflect the same trend as mentioned in Section A.1. See the following quotes, as examples:

- *"There wasn't much difference between the two subjects so I think that the non-white person would slightly get more benefit."*
- *"The way our court systems are set up most whites benefit a lot more than non-whites."*
- *"This subject is younger with less prior counts so will very much likely benefit from a better life."*
- *"Assuming gender worked against subject 1, he caught a break."*

In addition to these attributes, in part 2 we provided the participants with the true labels (i.e., whether each defendant will reoffend in the future). While they generally thought a participant who will not reoffend in the future is more deserving of leniency, some noted that such information is not available at the time of decision-making, so neither defendant should get a preferential treatment.

The following quotes from our participants are meant to show the diversity of opinions regarding the role of true outcome in determining deserved sentencing decision.

- *"With the benefit of knowing the actual outcome, it is easy to say that subject 2 deserves the more lenient decision."*
- *"You can be more lenient if they won't offend again."*
- *"Will not reoffend, so in theory, the system should be more likely to give them a more lenient decision."*

- *"Ultimately i believe the severity of the crime is the most important factor for the punishment besides that whether or not the reoffend would be second, but that is not possible to know ahead of time."*
- *"Hindsight is 20/20 so I had to choose the person which did not re-offend, however at the time of sentencing these two [defendants] would be utterly even."*
- *"since [subject] 1 is not white, I want to be lenient due to how minorities are put in prison at such large rates, but I had to consider if they will reoffend."*
- *"You don't have actual outcome making the decision, not sure if it belongs in the question here."*
- *"I'm ignoring the actual outcome because they both have the same data."*

A.3 Response to Part 3

The justifications provided based on the defendant's attributes, such as race, age, and prior counts, generally reflect the same trend as mentioned in Section A.1, indicating that according to our participants, historically disadvantaged groups generally benefit more from lenient decisions and are harmed more by harsh sentencing decisions. In addition to these attributes, in part 3 we provided the participants with the true labels and predicted labels (i.e. whether the algorithm predicts each defendant has a high risk of reoffending in the future). Pariticpants generally agreed that a low-risk prediction is more beneficial to the defendant, while a high-risk prediction is particularly harmful to the defendant if he/she does not have a long criminal history and is not going to reoffend.

The following quotes from our participants are meant to show the diversity of opinions regarding the role of the predicted label in determining deserved sentencing decision.

- *"Subject 2 was considered low risk to reoffend, so they would probably get a lesser sentence or no jail time. So they would benefit a lot more."*
- *"They were listed as low risk and were in fact low risk, so it helped them out a lot."*
- *"even with a felony charge, this subject still got a low risk from the algorithm so I think he benefited more from it's decision."*
- *"[Subject] 2 would benefit more in being given the benefit of the doubt when it was not actually warranted."*
- *"because subject 1 committed a felony and the algo concluded that subject 1 was at a low risk to reoffend. meanwhile subject 2 only committed a misdemeanor with all other attributes the same, yet the algo concluded that subject 2 was at a high risk to reoffend. so clearly subject 1 would benefit more from the algo's decision."*
- *"they are listed as low risk but will reoffend so it helped them personally as a criminal to be considered low risk."*
- *"if the algorithm says someone has a low risk to offend and they actually do reoffend, then putting back on the streets gives them an opportunity to do what they want to do."*
- *"subject 1 would benefit less because he has a lesser charge degree, but still gets the decision "high risk to reoffend.""*

B Additional Experiments

We repeated our experiments with a modified UI, in which we provided a neutral response option to participants. Figure 10 shows the weight vectors obtained through this UI using the responses of 20 AMT participants. We do not observe significant differences between the weight vectors learned through our original UI and the modified version.

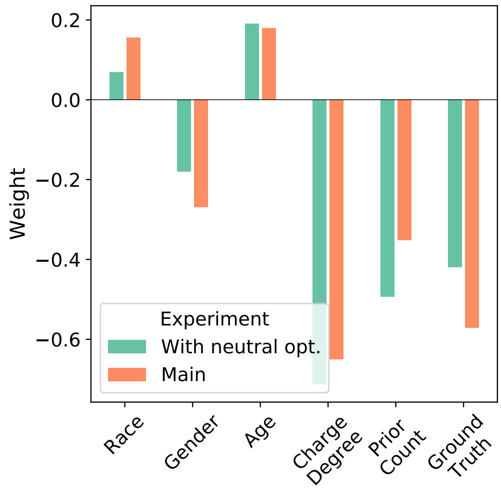

Figure 10: The weight vector, $\boldsymbol{\delta}$, estimated using the responses of 20 AMT participants through a modified UI with a neutral response option.

We also repeated our experiments with a modified version of our second questionnaire, in which participants were given access to defendants' predicted label (as well as their attributes and true labels). The weight vectors obtained using the responses of 20 AMT participants are displayed in Figure 11. We do not observe significant differences between the weight vectors learned through our original UI and the modified version.

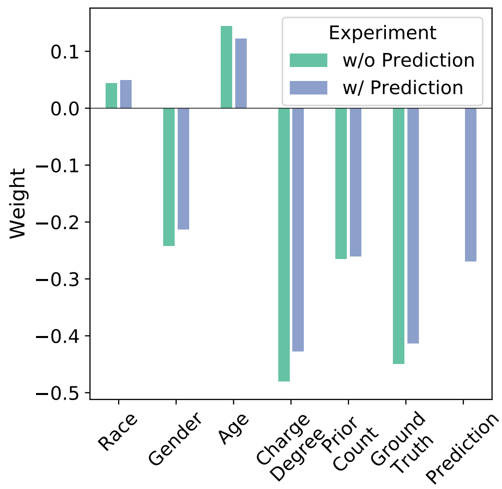

Figure 11: The weight vector, $\boldsymbol{\delta}$, estimated using the responses of 20 AMT participants—provided to a modified version of our second questionnaire, in which participants can observe defendants' predicted label. The log-likelihood of observing this data is -7.17. If we exclude $\hat{y}$, the likelihood reduces to -8.00.

C Long Introduction to the HIT

Figure 12 shows a longer version of the introduction to the HIT, including several examples.

Data-driven decision-making algorithms are increasingly employed to automate the process of making important decisions for humans, in areas such as credit lending, medicine, criminal justice, and beyond. **Organizations** in charge of decision-making can utilize massive datasets of historical records to learn a **decision-making rule** capable of making accurate **predictions** about never-before-seen individuals. Such predictions often serve as the basis for consequential **decisions** for these individuals. (From this point on we will refer to **individuals** subject to decisions as **decision subjects**.)

In recent years, several studies have shown that **automated/algorithmic decisions** made in the above fashion may disparately impact certain groups and individuals. For instance, in the context of credit lending, the decision-making rule may systematically disadvantage loan applicants belonging to a certain racial group and reject their loan applications more frequently. These observations have raised many questions and concerns about the fairness of automated decisions.

The goal of our study is to understand your moral reasoning and perception about what it means for automated decisions to be fair—considering the specifics of the decision-making context. We would like to know your ethical judgment through your answers to the following questions:

- *Which **attributes** of a decision subject do you consider **morally acceptable** for the decision-making rule to base its decisions on?*
  For instance, in the context of credit lending, one may consider it acceptable to for an individual's prior default history to affect the prediction he/she receives, but find it unacceptable for his/her parents' wealth to impact whether he/she is approved for a loan.
- *Comparing the attributes of two decision subjects, which one of them do you believe is more **deserving** of receiving **a better (more desirable) decision**?*
  For instance, let's assume all loan applicants prefer being approved to being rejected for a loan. Now consider two loan applicants with similar attributes, except for their number of prior defaults. One may consider the applicant with fewer prior defaults to be more deserving of being approved for a new loan.
- *How do you believe automated decisions will **impact** these subjects?* We would like you to imagine how algorithmic decisions may contribute to the **overall happiness, satisfaction, and well-being of a decision subject.**
  For instance, consider two loan applicants with similar attributes, except for their age and loan purpose—one being an 18-years old (pre-college) student who needs the loan to enter college and the other a 70-year-old who needs the loan to renovate parts of his house. In this hypothetical example, one may believe that the former decision subject (i.e., the student) would benefit more from receiving the loan.

This HIT consists of 3 parts. The first part contains 5 questions and the second and third parts each consist of 25 questions. Once you have completed the HIT, we will ask you to optionally provide us with some information about yourself. You will have the option to not respond to questions in this part. Your compensation for the HIT will not be impacted by how you respond to this exit questionnaire. For further details about our study, please read this information sheet.

Compensation and Bonus The basic compensation for completing this HIT is $6. In addition to this basic rate, you can earn a bonus of $2.5 by providing **thoughtful** justifications for some of your answers to the questions in the survey (you must provide justifications for at least 5 of your answers, that is, at least 1 justification for your answers to questions in part 1, 2 for part 2, and 2 for part 3, to qualify for the bonus). You will receive the basic compensation within 24 hours of successfully submitting the HIT (unless we detect that you have not paid sufficient attention to the task and your answers are low-quality/random, in which case we reserve the right to withhold compensation). The survey administrators need to review your justifications to decide if you qualify for the bonus. This process can take up to 10 days.

Figure 12: A more detailed version of the introduction to the HIT.

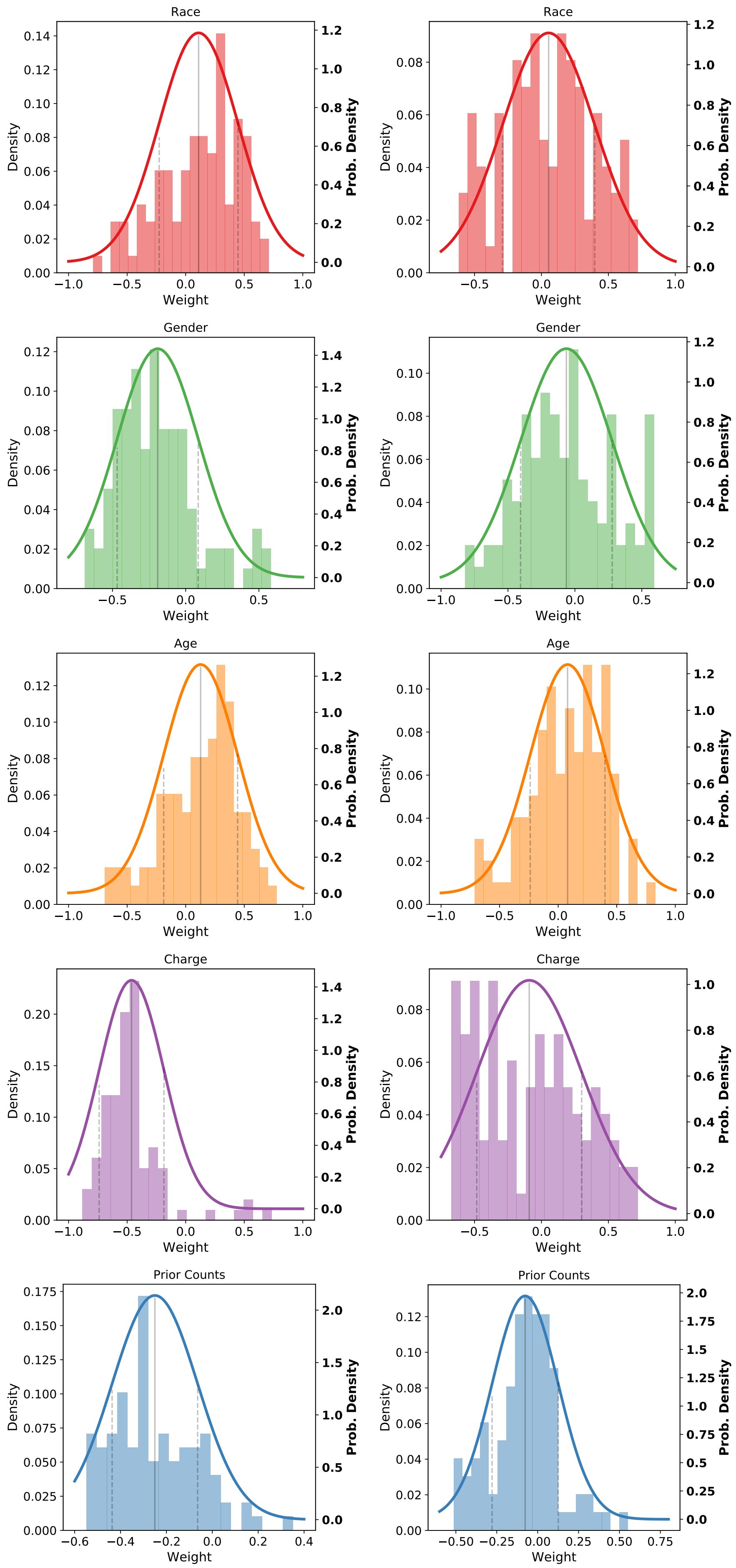

Figure 13: (Left) From top to bottom: the empirical distribution of δ on various features when $\boldsymbol{\delta}$ is estimated for each of our 99 participants, separately. (Right) The distribution of υ for the same setup.