TOWARDS UNDERSTANDING DECISIONS ABOUT FAIRNESS IN MACHINE LEARNING

A THESIS SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
MASTERS

DEPARTMENT OF COMPUTER SCIENCE

BY
GALEN HARRISON

CHICAGO, ILLINOIS
OCTOBER 30, 2019
For Rachel
# Table of Contents

ACKNOWLEDGMENTS .................................................. v

ABSTRACT ............................................................. vi

1. APPROACHES AND DEFINITIONS .................................. 1
   1.1 Mathematical approaches to Fairness ......................... 2
      1.1.1 Fairness as a property of the underlying distribution 3
      1.1.2 Using an impermissible attribute ....................... 5
   1.2 Tradeoffs and Decisions ...................................... 8

2. A RESEARCH AGENDA TOWARDS DESIGN DECISIONS ............... 11
   2.1 Data Science Workflow ........................................ 12
      2.1.1 Data science workflow tools ............................ 13
      2.1.2 Future work .............................................. 15
   2.2 Visualizing Fairness .......................................... 16
      2.2.1 Fairness Visualizations .................................. 16
      2.2.2 Future Work .............................................. 20
   2.3 Fairness and the public ...................................... 21
      2.3.1 Empirical Studies of Fairness ............................ 22
      2.3.2 Future Work .............................................. 24

3. AN EMPIRICAL STUDY OF DECISIONS ............................. 26
   3.1 Methodology .................................................. 28
      3.1.1 Survey Structure .......................................... 29
      3.1.2 Quantitative Analysis ................................... 32
      3.1.3 Qualitative Analysis .................................... 33
   3.2 Results ...................................................... 33
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.2.1</td>
<td>Participant Demographics</td>
<td>34</td>
</tr>
<tr>
<td>3.2.2</td>
<td>Preferred Trade-offs</td>
<td>35</td>
</tr>
<tr>
<td>3.2.3</td>
<td>Human Judges vs. Models</td>
<td>37</td>
</tr>
<tr>
<td>3.2.4</td>
<td>Fairness and Bias</td>
<td>39</td>
</tr>
<tr>
<td>3.2.5</td>
<td>The Impact of Who Was Disadvantaged</td>
<td>46</td>
</tr>
<tr>
<td>3.3</td>
<td>Discussion</td>
<td>47</td>
</tr>
<tr>
<td>3.3.1</td>
<td>Limitations</td>
<td>48</td>
</tr>
<tr>
<td>3.3.2</td>
<td>Comparison with other empirical studies</td>
<td>48</td>
</tr>
<tr>
<td>3.3.3</td>
<td>Conclusions and Future Work</td>
<td>50</td>
</tr>
</tbody>
</table>

REFERENCES : 52

A APPENDIX : 58

A.1 Additional Figures : 58
A.2 Demographic Models : 59
A.3 Survey Instrument : 59
A.4 Comparison of model visualizations with other empirical work : 69
ACKNOWLEDGMENTS

First and foremost I would like to thank Blase for agreeing to take this project on in the first place. Julia Hanson’s contributions, in the initial formation of this project, in writing the ConPro and FAT* submissions, and of course in actually developing the survey instrument were all key to the success of this project. In addition, Julio Ramirez and Christina Jacinto also deserve thanks for their enthusiasm and diligence this summer. Julia and Blase are both co-authors of the ConPro position paper from which the first and second chapters are adapted. Julia, Blase, Julio and Christina are also co-authors of the FAT* submission from which the third chapter is adapted. I also want to thank all members of the SUPERgroup, as well as Andrew, Will, and Nathan for their friendship.

More personally, my parents also deserve thanks for their guidance and support. And of course, I want to thank Rachel and Kubla. I literally could not have done this without you.
ABSTRACT

There are many competing definitions of what statistical properties make a machine learning model fair. Unfortunately, these definitions are in many cases are mutually exclusive, requiring a decision to favor one side of a tradeoff over the other. This is an example of a decision about a model. Current work in fair and transparent machine learning has neglected decisions up to this point. In particular, current data science workflow tools are not well suited to address identification and characterization of these decisions. Furthermore, decisions about a model are a key point of leverage for regulatory intervention as well as public engagement. It is not currently clear what the best practices are for visualizing these decisions in an unambiguous and non-manipulative form. This thesis describes a research agenda towards addressing these open questions. Furthermore, this thesis describes an empirical study of whether tradeoff decisions about a model are a sensible framework for understanding fairness for non-technical persons. Understanding perceptions of fairness tradeoff decisions, whether participants understood and had opinions about the tradeoff and if so what those opinions were, is necessary to form tools for public engagement. In the study, my coauthors and I had participants compare two models for deciding whether to grant bail to criminal defendants. The first model equalized one potentially desirable model property (with the other property varying across racial groups). The second model did the opposite. We observed a preference among participants for equalizing the false positive rate between groups over equalizing accuracy. Nonetheless, no preferences were overwhelming, and both sides of each trade-off we tested were strongly preferred by a non-trivial fraction of participants. These findings suggest that tradeoff decisions are a useful construct for understanding the fairness of machine learning models, and therefore provides support for the research agenda described in this thesis.
CHAPTER 1
APPROACHES AND DEFINITIONS

In recent years, the rise of artificial intelligence and big data has led to increased deployment of automated decision-making systems. Companies and governments have begun to rely on predictive models to make automated decisions about who gets things like jobs, loans, or bail. Automated decision systems seem attractive because they have the potential to make better decisions than a human decision maker. At the same time, because these sorts of decisions can seriously impact someone’s life, there are concerns about whether these systems can make these decisions ethically.

Consider the well-worn example of COMPAS. In 2016, ProPublica reported that COMPAS, a ML tool used to assess criminal defendants’ fitness for parole or bail, had a false positive rate for black defendants nearly twice that for white defendants [10]. The tool’s makers responded by noting that the accuracy of the tool was equalized between the two racial groups and that the tool was therefore not biased [22]. In response to incidents like this, researchers have developed several mathematical definitions of fairness. Each of these definitions try to capture slightly different aspects of the problem. Furthermore, just as certain philosophical notions of fairness are incompatible with other philosophical notions of fairness, these definitions also may exhibit incompatibilities or tradeoffs. For example, the COMPAS example spurred findings that one cannot simultaneously equalize accuracy and false positive rates across racial groups [17, 44].

Mathematical definitions can be used as tools towards achieving socially desirable outcomes, but do not in and of themselves capture “true fairness.” Definitions can be used to test for or prevent a particular unfair phenomenon, but it is the set of phenomena the model designers attend to which determines whether the model reflects a just outcome. It is not clear whether the makers of COMPAS were aware of the disparate false positive rates prior to ProPublica’s reporting, so it is not clear if preferring equalized accuracy was a conscious
decision. This ambiguity motivates this thesis’s call for further research into documentation of decisions in Section 2.1. Regardless, COMPAS implicitly embeds the notion that equalizing accuracy, rather than the false positive rate, is the correct definition of fairness. Not only did the makers of COMPAS make this decision entirely on their own, they did not even communicate that they were making such a decision.

This thesis is split into three parts. The first chapter summarizes prior approaches to fair machine learning, and provides an outline of the sorts of decisions about machine learning models which may be required. The second chapter describes an agenda towards addressing the lack of support for transparent and accountable decisions in machine learning. The last chapter describes a study I, and several of my collaborators carried out in order to test empirically whether decisions are a meaningful construct for understanding fairness outside of people who study issues of algorithmic justice. This thesis is adapted from a position paper presented at the FTC Workshop on Consumer Protection in May of 2019, and work submitted to FAT* in August of 2019 [36].

1.1 Mathematical approaches to Fairness

As noted by Mulligan et al., the term “fairness” as applied to machine learning has been used to refer to overlapping concepts [49]. Partially, this is due to the fact that fair machine learning is necessarily an interdisciplinary field, and different disciplines operationalize fairness in different ways. Partially, Mulligan et al. argue, this is due to the fact that fairness is an essentially contested concept; that is to say that the contestation of what fairness entails is an essential characteristic of fairness as a concept. While work within the fairness, accountability and transparency (FAT) community has generally analyzed fairness as a technical property - specifically how model inputs are mapped to outputs, other studies have noted that often concepts like control are subsumed under the banner of fairness [16]. The point is, what fairness is is not straightforward. Therefore, attempting to resolve what actually
makes a model fair is almost certain to be a hopeless task. Rather, it is more useful to consider the many technical and mathematical considerations of fairness as tools which may be deployed as a means of achieving an outcome which is considered just by all stakeholders.

Generally, fairness definitions have focused on particular aspects of model construction or behavior. These definitions imply different, possibly mutually exclusive, understandings of fairness [50]. Though overlapping in certain respects, these definitions can roughly be classified on the basis of where they identify the problem. Some identify unfairness as naturally arising from interactions between probability distributions. Another line of work identifies unfairness as consisting of decisions made on an impermissible basis (such as race or gender). Subsequent work has complicated the model-centric view of fairness by identifying a broader set of decisions about the model, such as the actual learning task and data filtering and how people interact with model decisions, which can also have fairness implications. The space of tradeoffs between these notions has not yet been entirely mapped out, as it is at least partially dependent upon the particular distributions underlying a problem, but it is known that at least some tradeoffs between these concepts exist.

1.1.1 Fairness as a property of the underlying distribution

Certain lines of work have tried to compensate for external conditions which induce unfair outcomes. In concurrent works, Chouldechova and Kleinberg et al. both consider situations in which one has an imperfect classifier and two distinct groups in which the true rate differs [17, 44]. Both Chouldechova and Kleinberg et al. note that in this case, it will be impossible to achieve a risk classification which achieves “test fairness” (equivalent for our purposes to accuracy), while also equalizing false positive rates. Insofar as these authors are actually proposing equalization of false positives as a definition of fairness (both authors take a critical stance towards ProPublica’s reporting on this issue), unequal false positives arise out of a difference in the base rates of recidivism between White and African-American populations.
Thus a characteristic of the underlying distribution leads to unfairness.

Another line of work has characterized unfair discrimination as arising as a natural consequence of greater uncertainty about smaller groups. This problem setting, sometimes called a multi-armed bandit setting, model may be more uncertain about predictions for a minority population than for the majority population, making the model less likely to take risks on the minority population [40]. Instead of binary classification, or risk prediction, the problem setting is one where the model can choose to sample from several different unknown probability distributions. At each point, it receives a payoff dictated by the distribution it chose. The model’s goal is to maximize the total payoff. Fairness, for this line of work, consists in achieving the global maximum. The intuition is that an unfair choice will choose a lower-payoff, but more well-known, distribution over a higher-payoff but less well-known distribution. In this framing, the problem is that a minority population is just by definition smaller than the majority population.

Work related to the multi-armed bandit problem has been applied to investigate the problem of censored feedback [26]. Censored feedback is the phenomena in which acting on a prediction affects data received in the future. For example, in predictive policing, an assessment of a neighborhood as high risk will lead to the assignment of police officers to that neighborhood. In turn, the increased officer presence will mean that crime in other neighborhoods will go unnoticed. Analogously to the multi-armed bandit setting, Elzayn et al. consider a decision to allocate these kinds of resources fair if it leads to an approximately equal chance of crime discovery across all groups.

These ways of thinking about where the problem arises is agnostic as to the source of the different distributions. Reading further into this, one can see these definitions as characterizing all involved parties as blameless. A public who picks up on these characterizations may, fairly or not, feel that these ways of understanding the problem fail to hold actual malefactors wrong, particularly if the differing underlying distributions are due to long histories.
of malfeasance. For example, the statement that increased police presence on the south side of Chicago is a natural consequence of selective feedback ignores deliberate patterns of deliberate discriminatory behavior towards this part of the city.

1.1.2 Using an impermissible attribute

Another line of work has rooted its analysis of fairness in the idea that fairness comes from making a decision for the right reasons. That is - if classifier makes classifications on the basis of an impermissible attribute (such as race or gender), then the classifications are unfair. The main issue this line of work struggles with is what it means for a model to be “using” a particular feature or not. The centralizing concern is that several non-sensitive features may be combined to act as a proxy variable for a sensitive feature. One example of this is the “fairness through awareness” advocated by Dwork et al. [25]. In this approach, disparity in outcomes between similar people should then be measured and limited. A key finding of Dwork et al. is that merely not including a variable will not prevent discriminatory outcomes. However, what constitutes similarity depends on the specific type and context of the data, and effectively determines the outcome. This requires a metric of similarity between individuals. For example, a credit score purports to measure a person’s creditworthiness, so one can compare two people’s creditworthiness by comparing their credit scores. Despite this approach’s insight about proxy variables, it has faced barriers to adoption because it is not clear how to measure similarity in any given situation.

Another example of an approach concerned with avoiding decisions made for the wrong reasons is the application of the legal notion of disparate impact to classification. Disparate impact is the legal concept, from employment discrimination law[^1] that a facially neutral test may be discriminatory if it excludes substantially more of one group than another. Disparate impact provides a way around the need to prove intent in all cases of employment discrim-

Applying a similar logic to classification decisions, Feldman et al. suggest that a classifier which disproportionately classifies favors one group indicates that the decisions are being made on an impermissible basis [27]. They show how disparate impact implies a bound on the statistical predictability of the sensitive attribute from the classification.

Hardt et al. also propose a solution to making decisions for the wrong reasons [34]. They propose two fairness definitions: in the first, false positives and true positives are equalized, in the second, just true positives are equalized. Though they describe this work as motivated by a desire to avoid decisions made on the basis of proxy variables, their approach suggests a concern more close to the line of thinking described in Section 1.1.1. In some ways, the actual definitions they propose are an attempt to find a nexus between the impossibility of equalizing false positives and accuracy. Supporting this interpretation, subsequent (which Hardt was also involved in) has identified scenarios in which definitions are satisfied, but which they argue have different fairness interpretations [43]. The differentiating factor Kilbertus et al. identify in this second work is the causal mechanism generating the data. In other words, the “why” matters in this new approach. However, because the causal mechanism matters, it means that applying this definition requires an understanding and model of the mechanism generating the data.

It is worth mentioning that several transparency mechanisms can also be seen as attempts to avoid models making decisions “for the wrong reasons”. Quantitative Input Influence (QII) quantifies the relative influence of particular variables over the output of a black-box model [21]. For example, QII can show how much race influences the model’s decision, and how that compares with the effects of gender. Another transparency tool is Locally Interpretable Model Explanations (LIME) [52]. LIME fits an interpretable model to the area directly around a particular data point. The interpretable model can then be used to explain how the more complicated model made its decisions. A full overview of explainable machine learning is out of the scope of this chapter, however I bring up these definitions in
order to highlight their connections to fairness approaches. Transparency as to the effects of particular variables is implicitly a statement about fairness. If impermissible attributes have large effects on the model’s predictions, then the model may not be considered fair.

What is a sensitive attribute?

The previously discussed approaches to fairness have all assumed the existence of a single sensitive variable, such as race, or gender, against which the definitions may be tested. However determining what variables are sensitive or not is not necessarily straightforward. Heidari et al. have attempted to analyze various forms of fairness into a Rawlsian notion of equality of opportunity [37]. In this they distinguish between variables over which the data subject (the person subject to the classification decision) has control, and variables over which the subject has no control. They then argue that fairness consists of all people having a roughly equal chance of receiving an unearned benefit. Other work along this line has argued that classifications should only be made on the basis of variables which the data subject can exercise control [57].

Grgić-Hlača et al. define fairness based on the process by which decisions are made, rather than the outcomes [32]. This notion of process fairness is achieved if the only predictive features used in a model are those that people believe fair to use, documented empirically via a survey. The survey asks if the respondent deems a particular feature fair to use, fair to use if it increases disparity, and fair to use if it increases accuracy. The authors find that one can sometimes achieve both process fairness and other more outcome-based fairness, but at an accuracy cost. Process fairness departs from other definitions by explicitly considering input from the general public. Follow-up work seeks to understand why particular features are considered fair or unfair [32].
1.2 Tradeoffs and Decisions

There are a wide range of fairness definitions, encoding a variety of understandings. Fairness can encompass a wide range of technical and social properties. Thus, a data scientist who wants to make their model fair will be faced with a significant set of decisions. Unfortunately, certain key fairness definitions are mutually exclusive or otherwise incompatible. Kleinberg et al. [44] and Chouldechova et al. [17] both independently argue that it is only possible to achieve equally accurate risk scores across groups and equally balanced risk quantiles across groups under very specific conditions. Grgić-Hlača et al. find that one can sometimes achieve both process fairness and outcome fairness, but at an accuracy cost [32]. Furthermore, the use of race may be necessary during model development to audit the model or to achieve group fairness [59]. Indeed, the lack of data labeled with unfair attributes can prevent the analysis of fairness in practice [38]. Friedler et. al. conducted a benchmark of various fairness tools and measures, and found that there are significant differences in accuracy and fairness measure depending on the preprocessing and method being applied to the data set [28]. These are all examples of tradeoffs just within fairness definitions.

Having selected a definition or definitions, the data scientist also needs to determine how strictly the model must adhere to the definitions. Many, if not all, of the algorithmic tools available to achieve a particular fairness definition fulfill the definitions only approximately. In many cases, the closeness of the result to a definition is configurable. For example, the disparate impact measure used by Feldman et. al. has a parameter $\tau$, which controls the acceptable difference between positive selection for the majority group and positive selection for the minority group [27]. It is suggested that $\tau$ be set to be 80% to be in line with the Equal Employment Opportunity Commission’s guidance on disparate impact claims, however a data scientist who sets it at that level may still create an unfair outcome. Courts have found instances where differences of 1% were sufficient to create a claim under employment anti-discrimination law [6].
Even in instances where the adherence is not directly configured, the data scientist may choose the degree of adherence inadvertently through their pre-processing choices. A data scientist evaluating a model for fairness needs to decide on more than a metric. In the preceding section, I have discussed the idea of fairness with respect to some group or variable. Despite gesturing at some possible variables or groups, I have discussed these in abstract terms. This reflects the way in which these definitions are framed and discussed. However, a data scientist seeking to make a model fair will need to make decisions about which sorts of groups are important. Both ProPublica and NorthPointe agreed that race was a meaningful lens to examine COMPAS through. A data scientist who evaluates a model for fairness with respect to race but not income or disability implicitly makes a determination that race is a salient concern when disability is not. To be clear, this assumption may be warranted, but the assumption should be made clear.

While the definitions chosen constitute a highly visible example of how problem construction can influence fairness decisions, it is by no means the only one. Mitchell et. al. discuss other aspects of the problem construction which can impact the fairness of a model [48]. The learning task, as well as the set of outcomes can also affect the fairness of the model. For example, a model built with the purpose of allocating police patrols is different from a model built to allocate social workers. The way success is measured may also impact how the model is received. The COMPAS model measured success by whether or not the candidate for parole recidivated, but it might also reasonably tried to minimize overall disruption to the community to which the person would be released.

Even beyond the strictly technical aspects, “fairness” also can be affected by interactions between various human actors and the model [8]. A series of studies have found that people often only follow algorithmic suggestions when it confirms their prior prejudices [29, 51].

The particular decisions confronting a data scientist will vary based both on the problem and available data. Currently, data scientists make these decisions, but in many cases fail
to make these decisions explicit. There are very few mechanisms for helping data scientists
navigate these decisions, even if they want to be more open about them. Furthermore, there
is little research into how the public actually thinks these decisions should be made. In the
next chapter, I will describe a research agenda for addressing this gap. In the chapter after
that, I will present a study into public attitudes towards fair decision-making.
CHAPTER 2

A RESEARCH AGENDA TOWARDS DESIGN DECISIONS

In practice, the data scientist building the model necessarily must make difficult tradeoffs choosing between imperfect models, balancing different definitions of fairness with accuracy and other considerations. Because these choices have ethical dimensions, there is a need to better support these choices and both document and justify them for the public. A data scientist in pursuit of a ‘fair’ model must make a series of key decisions about how to define fairness and how to balance fairness with other considerations. Further, a data scientist must weigh the model’s fairness against its other properties, like its overall accuracy or computational cost. However, current data science workflow tools do not adequately keep track of or characterize these decisions. In order to avoid data scientists making decisions without realizing it, further work into tracking the data science process and identifying and characterizing decisions is needed.

A data scientist should be aware of the tradeoffs they are making and also be able to clearly defend and communicate these decisions. Existing tools for providing fairness explanations in machine learning models focus on enumerating how a particular model makes decisions (Section 2.2), rather than highlighting nuanced comparisons between competing models. Visualizing and communicating decisions is necessary for a plurality of reasons. It is necessary for the data scientist to understand the decision, when confronted with one. It is necessary for the public to understand the decisions that were made in developing a model, in order to make informed choices about whether the model acts in the public interest.

Running through both of these themes is a concern with interactions between the model building process and various human actors. Empirically studying attitudes and perceptions of these tradeoffs is also necessary. Understanding currently existing attitudes towards decisions will help validate decisions as an operationalizable construct for deliberation about data driven decision making. Decisions may require a significant amount of translation to
make them legible to the non-technical public, which would suggest that decisions may not be the most useful for helping non-technical publics deliberate about their adoption. Luckily, this is not the case, as will be discussed further in Chapter 3. Furthermore, it will also be necessary to understand how different ways of visualizing and characterizing decisions produce different understandings in the public. Incorporating public reactions into the building process is useful, however it is not currently clear how to do this. Simply crowdsourcing decisions is not likely to actually produce an acceptable model.

This chapter describes the current needs for research into the following domains:

1. Creating data science workflow tools to track and characterize decisions.
2. Creating information visualizations that compare competing models by fairness and other considerations.
3. Conducting empirical studies to understand perceptions of tough tradeoffs made in model selection.

2.1 Data Science Workflow

Even when fairness is not a design concern, model building is an iterative process. A typical data science workflow consists of four main phases: preparation of data, analysis and reflection of outputs (of data analysis scripts), and dissemination of results including the model. When using data science to create a predictive model, a data scientist cycles through analysis and reflection phases in a process of trial and error. They create multiple models from prepared data, compare their outputs, then tweak or create new models according to their findings, and repeat. This tweaking may involve sampling the data in a different way, tuning parameters, or changing the model architecture or loss function. However, in practice these steps often blend together.
These choices may reflect engineering (computational costs) or statistical (accuracy) considerations. However, when the model is being used for a socially significant function, fairness may also be critical. Despite the importance and nuance of these tradeoffs, descriptions of these choices and how they were made are rarely communicated. In particular, data scientists may not even be aware they were making a choice. Recall from the example of COMPAS, we have no idea what actual choices COMPAS was faced with or considered. How and when did Northpointe (the makers of COMPAS) choose their definition of fairness, and what alternatives did they consider? We know from Northpointe’s rebuttal that the accuracy of their model was the same between racial groups. Was this a conscious, principled choice, or was it a post-hoc justification? Descriptions of, or comparisons to, alternative models are absent from their report. Current workflow tools, while helpful, do not directly support the tracking or identification of analysis decisions.

2.1.1 Data science workflow tools

As machine learning has been increasingly integrated into products and business practices, researchers have started to pay attention to assisting and documenting aspects of the data science workflow. The motivation for these tools has been to streamline the process of building a model, either by minimizing the amount of redundant work, or by ensuring that the data product was achieved through a consistent process. However, these tools and approaches, where they are concerned with fairness at all, largely leave the hard decisions hidden.

Data science can be procedurally messy. Despite many idealized descriptions of the data science process as being composed of a series of distinct steps, a recent study by Alspaugh et al. found that often exploratorium overlaps with other steps in the process\[7\]. Furthermore, Alspaugh et al. find that data science is rarely driven by specific testable hypotheses. Instead the hypotheses are developed and refined through the process of analyzing the data. A
consequence of this is that there is a need by data scientists for documentation and notes on prior analyses. Indeed, other recent work evaluating a system for identifying past analysis choices in Jupyter notebooks has shown promising results [42]. Being able to audit past analysis choices is useful and a necessary component of tracking and identifying decisions. However, an analysis choice is not identical to a decision. A decision is choice between options, whereas the analysis choices studied by Kery et al. are not necessarily concerned with this kind of counterfactual.

A further recognition of the need more broadly for documentation and tracking of analysis choices can be seen by the proliferation of various industry tools for tracking workflow, or at least past analysis choices. In 2016, Facebook announced that it had a framework to centralize past experiments and features used across by various internal machine learning teams [5]. Furthermore, the proliferation of startups like MLFlow and Databricks which specifically address the challenge of centralized documentation and logging of analyses suggests a serious need, separate from any fairness considerations, for documentation of process [60, 2]. Centralized provenance information and documentation is useful in identifying decisions, however deriving decisions from this kind of documentation requires significant interpretive effort.

In contrast to “bottom-up” tools, where the goal is to keep track of the analysis path as determined by the data scientists, there have been other efforts to impose process from the top down. These tools and approaches have sought to make the data science process more orderly, by imposing procedural, and sometimes computational restraints. Perhaps the best example of procedural constraints is the CRISP-DM framework. CRISP-DM, or the Cross Industry Standard Process for Data Mining, sets out six distinct steps and various questions to ask at each step during a data mining process [4]. These steps are meant to ensure that the ultimate product is useful to the business. While IBM does produce a tool meant to support people seeking to follow the CRISP-DM framework, the controls on this
process are largely procedural. In contrast, other work has proposed building a workflow directly into the database itself. The FIDES system is meant to ensure that discoveries are both statistically valid, and do not reflect any sorts of spurious correlations. Procedure, while it can help organize and surface them, is independent of decisions. In some cases, it may even hide decisions.

Neither the bottom-up approach, nor the top-down are particularly appropriate for the purposes laid out here. Ultimately the goal should be to identify and characterize decisions. Prior analyses, while components of a decision, are more granular than a decision. Reconstructing decisions from prior analyses may require significant interpretation and analysis. At the same time, the top-down approaches may constrain data scientists. Furthermore, while following a process may generate certain decisions, it will not by itself document or characterize those decisions.

2.1.2 Future work

Documenting the decisions involved in a particular model is crucial for transparency. Even if affected individuals receive an explanation of why the model made a given decision the way it did, they may very well ask why they were subject to a decision by that model, and not an alternative model, in the first place. Without an explanation of why this specific model was used, the picture provided to the public will necessarily be incomplete.

In order to achieve this, it is necessary to first understand and characterize the decisions the data scientists were confronted with, and what they chose. Currently existing support for documentation of machine learning processes does not directly address the issue of machine learning decisions. We propose that future work should seek to build systems integrations for tracking and characterizing decisions about a model. There is a need for research into tools which track analyses and which identify and characterize decisions.
2.2 Visualizing Fairness

Given adequate documentation about model decisions, the next question is how to visualize and communicate these decisions. In order to fulfill this role, the documentation must be understandable by a non-technical public. If the fairness decisions the data scientist made are obscured, the only way to convince a data subject that the model is actually correct is by appealing to the data scientist’s wisdom and knowledge. This is likely to be unconvincing. Alternatively, the data scientist may have misjudged the correct course of action. Clear, publicly available, decision documentation will assist in discovering and correcting these misjudgments.

To reason about choices Northpointe made, we argue that data scientists, regulators, and consumers ought to have a clear understanding of the tradeoffs between alternative models and an explanation of how tough choices were made about them. Further, understanding society’s beliefs about how these decisions should be made in different contexts will enable data scientists to make better decisions about models, while establishing best practices for documenting these choices could enable a consumer protection framework for fairness.

Communicating these choices, however, is difficult. Even the relatively constrained scenario described in figure 2.1 is already somewhat complex. There are eight discrete values to keep track of, some of which are comparable to one another, some of which are not. We have not included other arguably pertinent values like false negative rates. We also have limited the number of groups to two and the pertinent category to race. In reality, a data scientist will likely have more than two discrete options. This is why considering and communicating alternatives poses a significant research challenge.

2.2.1 Fairness Visualizations

Though recent attention to fairness has prompted the creation of several different methods of visualizing models, these visualizations are not well-suited to characterizing decisions. An
Figure 2.1: A hypothetical decision the makers of COMPAS may have been faced with when deciding which model to use.

Figure 2.2: Google’s What-If tool visualizes experiments on machine learning models, aiding in model selection.
Proportional Parity: Failed

<table>
<thead>
<tr>
<th>What is it?</th>
<th>When does it matter?</th>
<th>Which groups failed the audit:</th>
</tr>
</thead>
</table>
| This criteria considers an attribute to have proportional parity if every group is represented proportionally to their share of the population. For example, if race with possible values of white, black, other being 50%, 30%, 20% of the population respectively, has proportional parity, it implies that all three races are represented in the same proportions (50%, 30%, 20%) in the selected set. | If your desired outcome is to intervene proportionally on people from all races, then you care about this criterion. | For education (with reference group as Masters)  
Doctorate with 0.58X Disparity  
Bachelors with 1.30X Disparity  
10th with 2.33X Disparity  
11th with 2.29X Disparity  
9th with 2.32X Disparity  
Prof-school with 0.59X Disparity  
Assoc-voc with 1.76X Disparity  
Some-college with 1.95X Disparity  
7th-8th with 2.33X Disparity  
5th-6th with 2.34X Disparity  
Assoc-acdm with 1.75X Disparity  
1st-4th with 2.26X Disparity  
HS-grad with 2.06X Disparity  
12th with 2.29X Disparity  
Preschool with 2.37X Disparity  
For gender (with reference group as Male)  
Female with 1.29X Disparity |

Figure 2.3: A portion of a report from the Aequitas tool on the fairness of a given model.

2. Check bias metrics
   
   Dataset: Compas (ProPublica recidivism)  
   Mitigation: none

   Protected Attribute: Sex  
   Privileged Group: Female, Unprivileged Group: Male  
   Accuracy with no mitigation applied is 66%  
   With default thresholds, bias against unprivileged group detected in 4 out of 5 metrics

   ![Statistical Parity Difference](image1)  
   ![Equal Opportunity Difference](image2)  
   ![Average Odds Difference](image3)  
   ![Disparate Impact](image4)  
   ![Theil Index](image5)

Figure 2.4: The IBM AI Fairness 360 tool at the bias detection stage.

Example of a fairness visualization aimed at data scientists is Google’s What-If tool [3], shown in Figure 2.2. The tool facilitates exploratory analysis in model selection, highlighting the impacts of tuning hyperparameters and excluding particular variables on model accuracy. Its visualizations of how these considerations impact different demographic groups can be used to analyze fairness in part. Because it is an exploratory tool, it does not attempt to
communicate any sort of set of alternatives, nor is it designed to engage non-expert users.

In contrast, the Aequitas tool focuses on reporting the biases of a model [1]. Given a single model, Aequitas runs tests for a set of fairness definitions like those discussed in Section 1.1. While more straightforward for a non-technical person to understand (see Figure 2.3), Aequitas reports only on a single model. An Aequitas report does not clarify why the data scientists who built the model made the decisions that led to particular outcomes. Our vision instead requires that future work develop expanded tools for comparing among many models and documenting the difficult decisions made.

Last year, IBM launched AI Fairness 360 (AIF360), an open source collection of tools that, like Aequitas, runs tests for fairness metrics, but also provides a suite of possible algorithms for mitigating bias [13]. Along with the toolkit code, the AIF360 website hosts interactive demos of the tools in practice (see figure 2.4). We observe that the comparisons presented in these demos are limited to pairwise before and afters of models after a single mitigation step; comparisons between mitigation steps are absent. Current visualizations of fairness do not aim to communicate process or alternatives, both key components of a decision.

Usability of Visualizations

While the visualizations which explicitly engage fairness have yet to be studied in a rigorous fashion, some human-factors research has been applied to transparent methods discussed in Section 1.1.2. Ribeiro et al. conduct a limited user study in which participants use LIME to choose between two models. However, the decision is between a model that generalizes well and one that does not; no ethical choice is involved. Binns et al. examine whether the type of explanation a person receives about a decision affects how they perceive the justness of that decision [15]. They tested explanations based on QII, the amount the input would need to change to affect the output, the most similar case in the training data, and aggregate
statistics by category. They found that the style of the explanation only affected perceptions of justice when someone was exposed to multiple different explanations. Unfortunately, this work only examines explanations of individual predictions, and does not attempt to characterize the model.

Krause et al. try to bridge the gap between individual explanations and overall patterns by aggregating explanations provided by LIME \cite{45}. They validate their interface by testing whether participants can identify a deliberately biased model. Aggregating explanations may play a role in comparing two models, but is again insufficient on its own for justifying difficult decisions in model selection.

2.2.2 Future Work

Being able to effectively document and rationalize difficult choices in model selection will facilitate better communication with laypeople potentially affected by the model’s decisions. With visualization tools to explain differences between models, data scientists can solicit and use findings on public opinion about model fairness tradeoffs to justify their choices.

Model selection explanations could also be incorporated to improve classification explanations. For example, the plausible explanation “you were classified as a high insurance risk because you are a man, are between the ages of 18-22, and were born in April” communicates that gender, age, and birth month were the criteria used to determine his insurance risk, but fails to explain why those criteria were used. Including descriptions of the decisions made between model alternatives could make this explanation more satisfying. A better explanation could say, for example, that a data scientist chose to include birth month as a variable in their model because its inclusion equalized accuracy between men and women compared to other models that were otherwise similar.

While explanations can empower users, they can also mislead them. Lying with statistics, and in particular with statistical visualizations, is a well-known phenomenon. Data scientists
fearing backlash may seek to mislead the public to obscure or reframe the decisions they made in a positive light. Notably, researchers have found that most Facebook ad-targeting explanations are incomplete and vague [9].

A second line of investigation should be into techniques for presenting the decision in misleading ways. While heavily implicated in the visualization of the model qualities, this investigation needs to also look beyond the way the information is presented to the choice of information being presented. For example, one could imagine specifically choosing bad models to present next to the one ultimately chosen to take advantage of anchoring.

Lastly, investigating the role of affect and emotion in visualizing fairness is likely to be promising. Recently there have been calls for feminist data visualization which, among other things, more legitimize embodiment and affect [23]. Judgements about fairness and algorithmic justice tend to have some aspect of affect associated with them. Visualizations which more explicitly engage with affective aspects of fairness may more effectively communicate choices to the public.

### 2.3 Fairness and the public

Underlying the previous sections is a desire to foster public involvement in these decisions. Given that, a data scientist may want to consider what the public thinks of these tradeoffs. Once the data scientist has made their decision, they may want to communicate the rationale behind their decision and document the alternatives they considered. While we don’t propose that fairness should be crowdsourced, the public or publics should be have at minimum some form of meaningful transparency as to the decisions. We do not make any sort of specific claims as to how that transparency should be made meaningful, however we do note that a robust literature exists around design for the public (see e.g. [19]). Engaging the public in these decisions requires an empirical study of perceptions of model decisions. It doesn’t make sense to design for an audience which does not exist, or which will not understand the
decision being posed. While there is a burgeoning body of work studying fairness perceptions in algorithmic systems, these studies do not investigate attitudes towards decisions.

2.3.1 Empirical Studies of Fairness

Empirical studies of fairness, suggest that decisions may be a meaningful to non-technical audiences, but have not investigated this question directly. A handful of empirical studies have collected attitudes about specific model properties. In a loan allocation context, Saxena et al. quantify attitudes about individual fairness [53]. They ask participants to rate the fairness of three different ways a loan officer could divide $50,000 between two individuals with different repayment rates, and in one iteration, different races. Participants rated giving the entire sum to the candidate with the higher repayment rate as more fair than dividing it equally only when the candidate with the higher rate was black. Kennedy et al. investigate the relationship between trust and model properties. They use an experiment in which participants choose between pairs of risk assessment models [41]. The models vary at random in overall rates (e.g., true/false positives/negatives), the size of the training data, the number of features, the weight of features, the algorithm source, and more. While they investigate trust, they do not investigate fairness or test whether differences in model properties across groups affect trust. In a series of participatory design workshops, Woodruff et al. asked participants specifically drawn from marginalized communities about their experiences of algorithmic unfairness [58]. They found that participants did in fact have experiences with and reactions to algorithmic discrimination, though they did not automatically label their specific experiences as being the result of algorithmic processes. Another qualitative study examined reactions to an automated classroom team assignment tool and found that both the teachers who used the tool and the students to whom the tool was applied wanted greater control over the criteria the tool used [39]. While these studies do not directly engage with the question of decisions, they suggest that decisions may be a meaningful to people outside
of specialized technical communities.

Several studies have attempted to “crowd-source” fairness decisions. Srivastava et al. asked participants to choose between a succession of pairs of models to identify which group fairness definition best captures people’s perceptions of fairness \(^{55}\). Through twenty comparisons generated through an adaptive algorithm, they converge upon a given respondent’s preferred notion of fairness. In both risk prediction and medical diagnostic contexts, their participants prefer demographic parity (equality across groups in the percentage predicted to receive a positive classification) over other definitions. The respondents were asked to make twenty comparisons generated through an adaptive algorithm meant to converge upon the persons preferred notion of fairness.

Regardless, we note that the examples generated under this methodology do not necessarily encode any notion of trade-off between properties. For example, given the choice between 5\% greater accuracy overall (and higher inter-group disparity) or less inter-group disparity (and lower accuracy), which does the public find more fair? One study, which does purport to measure attitudes towards some form of tradeoff is the “moral machines” study by Awad et al. \(^{12}\). In this study, Awad et al. used a viral game to study variations across cultures in how participants would solve a variation of the trolley problem involving different demographic characteristics of the parties. The authors motivate this study with the goal of identifying universal principles for autonomous vehicles. Moral machines is instructive, though perhaps more as a negative example. One issue with moral machines is that it motivates the sorts of decisions it studies as necessary and inevitable, when in fact deploying autonomous vehicles is not at all necessary. In Chapter 3, we will present findings which suggest that these decisions are contentious and that relying upon crowdsourcing will not produce a widely acceptable model.

Grgić-Hlača et al. (discussed in Section 1.1.2) investigate attitudes about what features people consider fair to use \(^{31}\). In follow-up work, Grgić-Hlača et al. investigate why people
consider those features fair or not [30]. They find a feature’s perceived relevance, reliability, and volitionality drive assessments. They also find support for the explicit consideration of race increases from 17% to 42% when participants are told the use of race increases accuracy. This study is again suggestive of the sorts of considerations likely to be salient when characterizing decisions. Berkel et al. study the effects of deliberating on these sorts of judgments [14]. They effectively repeated Grgić-Hlača et al.’s experiment, but varied whether participants were placed into an anonymous chat room to discuss the choice. They found that more diverse groups reached decisions which were more unanimous after discussion than more uniform groups.

These studies suggest that decisions are meaningful to the non-technical public. Furthermore, they suggest that comparisons of accuracy and variable usage may be quantities of particular interest. Understanding public perception of machine learning decisions is not the same as determining a decision with data. The public writ large may have biases and prejudices that indicate an ethically impermissible course of action. Additionally, the data scientist may not have effectively communicated the tradeoffs. Furthermore, the correct decision may be contested - that is there may be multiple perspectives on the correct course of action.

2.3.2 Future Work

The three research strands discussed in this chapter are interrelated. Tracking the process in order to identify decisions is required in order to visualize decisions. How decisions are visualized impacts the accessibility to public or to study participants. Running through this is a concern with how the public will respond to and understand decisions. Because these are mutually reinforcing, the specific questions will likely evolve as progress is made on each aspect.

Our call to understand public attitudes towards model decisions begs the question do
people care? If they care, do they understand? It turns out that people do care, and at least to a certain extent understand. The study showing this is discussed in Chapter 3.

Beyond the study, understanding how visualization of the decision impacts public understanding is a key concern. Other open questions include a comparison of already-existing visualization techniques on fairness perceptions. The study described in Chapter 3, Krause et al. [45], and Srivastava et al. [55] each used distinct visualizations for comparing models. We found slightly different results from Srivastava et al. (Krause et al. did not ask a comparable research question), and think that the difference is likely due to their manner of visualizing the models.

Another compelling line of research is on techniques for incorporating public opinion into the design process. We do not think that merely crowd-sourcing the decision will be productive, so what is the best way to ensure that the machine learning system is a reflection of the public interest? There is a significant literature on community base design, largely focusing on issues of urban planning and community [24, 20, 11, 18, 47]. However, to our knowledge, none of these have considered issues of designing machine learning. Data driven artifacts likely pose new challenges for design because the full set of affordances may be less obvious to both the designer and to the community.
CHAPTER 3
AN EMPIRICAL STUDY OF DECISIONS

Few studies investigate attitudes about the difficult choices and fairness-related trade-offs inherent in realistic applications of ML. Thus far, the argument that decisions are a useful lens through which to understand models has relied on arguments based on research within a fairly specialized field. If the ultimate goal is greater accountability over decisions, it is reasonable to ask whether decisions have salience to the public outside of a highly specialized technical field. Understanding attitudes about these trade-offs can inform technical and regulatory interventions.

To gauge perceptions of the fairness of such realistic, imperfect models, my coauthors and I conducted a between-subjects experiment with 502 Mechanical Turk workers. We encoded this trade-off in a pair of models presented as straightforward bar graphs. Each pair captured one side of the type of trade-off that an ML developer might face when trying to make a model as fair as possible in realistic circumstances. For example, one set of participants was prompted to choose between a model that equalized accuracy at the expense of disparate false positive rates across racial groups, and one that did the opposite. We tested trade-offs of all pairwise combination of four desirable fairness properties: equalized accuracy; equalized false positive rates; equalized outcomes; and explicitly excluding race as a model input. Each model in a pair had a disparity in one model property and equality in the other. We also varied which racial group the model disadvantaged. Our results highlight challenges for building machine learning models that are perceived as fair and broadly acceptable in realistic situations. Despite this, they indicate an interest in and engagement with decisions about models.

Through this study, we answered the following research questions:

- **RQ1** When choosing between models exhibiting the two sides of a difficult trade-off, which do people prioritize?
We observed a statistically significant trend, but not full consensus, in participants prioritizing the equalization of false positive rates across groups rather than the equalization of accuracy. Notably, this preference is the opposite of what the COMPAS tool did, possibly explaining some of the controversy around the tool. For many other trade-offs, we saw a wide distribution in participants’ preferences. We observed a non-trivial fraction of strong opinions preferring each side of each trade-off we tested, highlighting the difficulty of building ML models with broadly acceptable fairness characteristics in realistic circumstances.

- **RQ2** What models that encapsulate difficult, yet realistic, trade-offs do people perceive as fair or biased?

Participants tended toward rating most models we tested as not biased. However, participants tended to consider it biased when outcomes were equalized at the expense of disparate false positive rates that disadvantaged African-American defendants. Only one model we tested — one in which more African-American defendants than White defendants were granted bail even though race was not used as an input to the model — was overwhelmingly considered fair. In many other cases, opinions were mixed and fairly polarized.

A model’s lack of bias did not necessarily imply its fairness. We observed a number of cases in which participants considered a model not to be biased, yet also did not consider it to be fair. In some cases, this was because of high false positive rates that were nonetheless equal across racial groups.

- **RQ3** Do people prefer to use an imperfect model or rely on a human judge?

For most trade-offs we investigated, we found a preference for a human judge over either ML model the participant saw. Even when the participant considered one or both of these models unbiased, they often still preferred a human judge. They justified this preference based on a human judge’s accountability, capacity for ethical reasoning, and ability to make individualized decisions.
• **RQ4** To what extent do responses vary based on which racial group the model disadvantages?

For each trade-off, we randomly assigned whether White or African-American defendants would be disadvantaged by the disparate property. We observed some potential, yet not statistically significant, differences in the distribution of responses.

In sum, our empirical user study is a first step in unpacking how people view the bias and fairness of ML models encoding difficult trade-offs related to fairness.

### 3.1 Methodology

To investigate perceptions of fairness in imperfect, yet realistic, ML models, we conducted an online, between-subjects, survey-based experiment. We asked participants to rate the fairness, bias, and utility of two models that exhibited both sides of a specified trade-off between two fairness-related model properties randomly selected from among the following four: accuracy, false positive rate, outcomes, and the explicit consideration of race. We graphically presented the properties encapsulated in this trade-off. Participants then chose which of the two models they preferred overall, as well as whether they preferred a human judge to either ML model.

On Amazon’s Mechanical Turk, we recruited workers 18+ years old and located within the United States. We limited recruitment to workers with a 95%+ rating over 500+ HITs. We paid $2.50 for the survey, which took a median time of 14 minutes. We excluded data from participants whose free responses were off-topic or nonsensical (e.g., discussing the Tesla Model X car in their response). Our Institutional Review Board approved this experiment.

Below, we detail the structure of the survey (Sec. 3.1.1) and the specific trade-offs we tested (Sec. 3.1.1). We then describe our quantitative (Sec. 3.1.2) and qualitative (Sec. 3.1.3) analyses. The supplementary appendix includes the full survey instrument.
3.1.1 Survey Structure

To familiarize participants with the topic of machine learning, the survey began by giving a high-level description of how ML models can be used to make predictions. We then told them a city was considering using an ML model to decide whether to grant bail to defendants charged with non-violent crimes.

Each participant was randomly assigned a pair of models exemplifying a trade-off between two fairness-related model properties (see Sec. 3.1.1). Each model satisfied one definition of fairness, but violated the other. We presented these properties visually (see Fig. 3.1).

We first presented the models individually, in randomized order. Participants rated the fairness, bias, and utility of each on a five-point Likert scale. Afterwards, we presented the two models side-by-side (as in Figure 3.1), asking participants to rate on five-point Likert scales which was more fair, biased, or useful. Figure 3.1 shows an example of how we presented the pair of models. Also on five-point Likert scales, we asked participants to select
Table 3.1: The properties investigated. The equal accuracy rate, as well as disparate rates for outcomes and false positives, were based on the COMPAS dataset [46].

which model they would prefer to see implemented, as well as whether they would prefer a human judge to their choice of either model.

Model Properties and Trade-offs

Each participant evaluated and compared two models that were imperfect in opposite ways. We randomly assigned each participant a pair of models representing the trade-off between two of the following four properties: accuracy, false positives, outcomes, and explicit race usage. We chose these properties because they have been widely discussed and can be expressed succinctly. In contrast, we chose not to test more complex definitions like equalized odds, which requires equalizing both the true positive and false positive rates [35]. Table 3.1 summarizes these properties, and their associated disparate and equalized rates. We tried to capture realistic trade-offs by using rates taken from ProPublica’s analysis of COMPAS data [46]. Because ProPublica’s analysis does not quantify all of the properties we tested, we fabricated the unknown rates (disparate accuracy and equalized false positive and outcome rates).

Below, we list the properties (and, in italics, the abbreviations we use throughout the paper) as we explained them to participants:

- **Accuracy (Acc):** “The accuracy of the model is the rate at which the model makes correct predictions. A prediction is correct if the model either predicts that a defendant will show up for their trial and they would or that they will not show up for their trial
and they would not have.”

- **Outcomes:** “The probability of bail is the likelihood of a defendant being granted bail if the model is used.”

- **False positive rate (FP):** “A defendant who is mistakenly denied bail is one that the model predicts would not show up for their trial when they would have.” We explained false positives without using the term “false positives” to avoid confusion about what a “positive” classification meant.

- **Race usage:** If the model does not use race: “The model makes decisions with no knowledge of the race of the subjects. Other features like type of offense and number of previous offenses are used as input to the model.” If the model uses race: “The model makes decisions with knowledge of the race of the subjects. Other features like type of offense and number of previous offenses are used as input to the model.” We considered a model “equalized” if it did not use race as an explicit model input. We considered a model to be “disparate” if it did use race. This corresponds to the notion that considering race is generally undesirable.

For each participant, we randomly selected one of the six possible pairwise combinations of these four properties. In all models, we showed how the two properties varied across two subgroups: White defendants and African-American defendants. We also randomly assigned whether African-Americans were disadvantaged or whether White defendants were disadvantaged in the disparate rates, doubling the total number of conditions to twelve. We randomized the order in which the participant saw the models.

**Terminology**

As we present our results, we refer to the trade-off participants saw using the abbreviated names of the two properties involved, as well as the *disadvantaged* group. For example,
“FP-Outcome-Maj” refers to the trade-off between false positives and outcomes in which the majority group (White defendants) is disadvantaged. At some points, we need to refer to the particular model within the pair. We use = and ≠ to indicate the equalized and disparate property, respectively. For example, the FP-Outcome-Maj condition includes the model = Outcome, ≠ FP, Maj and the model ≠ Outcome, = FP, Maj. When we quote participants, we identify them by number and condition (trade-off).

3.1.2 Quantitative Analysis

We mapped answers on five-point Likert scales to (-2, -1, 0, 1, 2). For example, participants could rate whether they would definitely (-2) or probably (-1) prefer a model, that they were unsure (0), or that they would probably (1) or definitely (2) prefer a human judge. For such answers, we tested whether participants’ answers tended toward one answer (model), the other (human judge), or neither. We did so using the unpaired Wilcoxon Signed-Rank test, which tests whether a distribution is skewed around zero. Significance indicates answers tended toward one answer or the other.

For questions where participants individually rated each model (e.g., on fairness), we tested whether they tended to rate one model higher than the other. As each participant rated both models, the data was not independent. We thus used the paired Wilcoxon Signed-Rank test, which measures whether the distribution of differences between pairs of ratings is symmetric. Significance indicates one model was seen as more fair, biased, or useful than the other.

For each trade-off pair (e.g., FP-Outcomes), some participants saw a version where White defendants were disadvantaged (-Maj), while others saw a version where African-American defendants were disadvantaged (-Min). We compared the -Maj and the -Min versions of each pair using the Mann-Whitney U test (a non-parametric analogue of the ANOVA test for comparing two groups).
For each of the above families of tests, we corrected for multiple testing with the Benjamini-Hochberg method.

As both fairness and bias ratings were ordinal, Likert-scale data, we calculated the correlation of these ratings with Kendall’s $\tau$.

To test whether participants’ background and demographics correlated with the ratings they gave, we fitted Mixed-Effects Ordinal Regression models. We built models with fairness, bias, and usefulness ratings, respectively, as the dependent variable. We dummy coded race, education, technical experience, political affiliations as regressors. We included interaction terms between education and income, and education and technical experience.

3.1.3 Qualitative Analysis

We thematically coded free-response explanations participants gave for their choice of one model over another, their preference for a human judge over either model, and their ratings of fairness and bias. Two coders collaboratively developed a codebook from a sample of answers. Two coders then independently used that codebook to code the remaining answers. We allowed multiple codes per answer to capture the compositionality of responses.

We created three distinct codebooks. The first was for explanations of bias. It contained ten high-level codes, six of which had sub-codes. For this codebook, Cohen’s $\kappa = 0.77$. The second was for why a participant chose one model over the other. It had seven codes (no sub-codes), and $\kappa = 0.71$. The last codebook assessed explanations of why humans or models were preferred. It had eight high-level codes, five of which had subcodes. For this codebook, $\kappa = 0.64$. The coders met to resolve disagreements.

3.2 Results

We begin by describing our participants (Sec. 3.2.1). We detail which trade-offs participants preferred (Sec. 3.2.2) and whether they ultimately preferred a human judge (Sec. 3.2.3).
Figure 3.2: Participants’ preferences for one model within a trade-off pair over the other. The left and right arrows indicate preferences toward the side of the trade-off listed on the left or right axes of the graph, respectively.

We then unpack participants’ ratings of fairness and bias, as well as how those concepts correlate (Sec. 3.2.4). Finally, we delve into the impact of varying which racial group was disadvantaged (Sec. 3.2.5).

3.2.1 Participant Demographics

We surveyed 502 individuals of whom 59.6% self-identified as male and 39.6% as female. The remaining 0.8% of participants either declined to state their gender or selected gender non-binary. Our sample skewed young, with 48% reporting their age as 25-34 and 27% as 35-44. 41.6% of our pool reported an annual household income of $20,000 - $49,999, with the second most common category being $50,000-$99,999 at 34.4%. Political affiliation was split; 48% of our survey pool described themselves as Democrats, 28% as Independent, and 20% as Republican. Most (75.9%) of the respondent pool identified as non-Hispanic White. 10.2% identified as Black or African-American, 9.4% as Asian, and 5.2% as Hispanic or Latino. 42% of participants had a 4 year college degree, and 20% had some college. The rest split between high school graduate (16%), 2 year degree (13%), and professional degree (8%).
<table>
<thead>
<tr>
<th>Demographic</th>
<th>Our study</th>
<th>Redmiles et al.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>59.6%</td>
<td>50%</td>
</tr>
<tr>
<td>Female</td>
<td>39.6%</td>
<td>48%</td>
</tr>
<tr>
<td>White</td>
<td>75.9%</td>
<td>84%</td>
</tr>
<tr>
<td>Black or African-American</td>
<td>10.2%</td>
<td>10%</td>
</tr>
<tr>
<td>Hispanic/Latino</td>
<td>5.2%</td>
<td>4%</td>
</tr>
<tr>
<td>Other*</td>
<td>14.1%</td>
<td>5%</td>
</tr>
<tr>
<td>$0-$50,000*</td>
<td>52.0%</td>
<td>49%</td>
</tr>
<tr>
<td>$50,000 - $100,000*</td>
<td>34.4%</td>
<td>38%</td>
</tr>
<tr>
<td>$100,000+*</td>
<td>10.6%</td>
<td>11%</td>
</tr>
<tr>
<td>Less than High School</td>
<td>(n = 1)</td>
<td>0.4%</td>
</tr>
<tr>
<td>High School</td>
<td>15.9%</td>
<td>12%</td>
</tr>
<tr>
<td>Some college*</td>
<td>33.5%</td>
<td>41%</td>
</tr>
<tr>
<td>B.S. or above*</td>
<td>49.8%</td>
<td>46%</td>
</tr>
</tbody>
</table>

Table 3.2: The demographics of our participant pool, as compared with Redmiles et al. [51]. * denotes a retabulation of either our data bins or their data bins to have comparable binnings.

large portion had no experience with computer science (48%). A roughly similar percentage (41%) reported no experience with probability. Most respondents (77%) reported having no experience with ML.

At least with respect to gender, race, income and education, the demographics of our participant pool is typical of other MTurk studies. We provide a comparison of the demographic makeup of our participants with that of a study conducted in 2018 with comparable MTurk criteria [51].

### 3.2.2 Preferred Trade-offs

We investigated six pairs of trade-offs between properties, each with variants in which majority (White) and minority (African-American) groups were disadvantaged. We observed a preference for equalizing false positives over equalizing accuracy. As shown in Figure 3.2, this preference was statistically significant, yet not overwhelming. Comparing false positives against accuracy, 53.7% of participants probably or definitely preferred the model that equalized false positive rates when White defendants were disadvantaged, and 56.8% of par-
participants favored the model that equalized false positives when African-American defendants were disadvantaged. In contrast, only 14.6% and 18.2% of participants, respectively, probably or definitely preferred the model that equalized accuracy at the expense of disparate false positive rates. The rest were unsure.

That we observed this preference toward equalized false positive rates is particularly notable because ProPublica’s reporting centered on COMPAS doing the opposite: equalizing accuracy at the expense of disparate false positive rates [10]. In short, this result highlights that COMPAS equalized the property that our participants preferred significantly less, potentially explaining some of the public controversy about the COMPAS system.

Among the remaining five trade-offs, again shown in Figure 3.2, we also observed a statistically preference toward equalizing false positives over equalizing outcomes when White defendants were disadvantaged (FP-Outcome-Maj). Here, 47.4% of participants preferred the model that equalized false positives, while only 15.8% preferred the model that equalized outcomes. However, this effect was not statistically significant when African-Americans were disadvantaged, nor were preferences statistically significant for any of the other four trade-offs.

Notably, for each trade-off, a non-trivial fraction of participants preferred each side of the trade-off (see Fig. 3.2). Recall that participants were assigned to one of twelve conditions (six trade-off pairs multiplied by two different groups that could be disadvantaged). For the nine conditions in which we did not observe a statistically significant preference, at least 24.4% of participants probably or definitely preferred each side of the trade-off. Furthermore, at least 8.9% of participants definitely preferred each side of the trade-off. In other words, a non-trivial fraction of participants would be unhappy no matter which side of the trade-off was chosen.

Our qualitative analysis of participants’ free-text justifications for their choice emphasized that participants successfully identified the characteristics we were varying, yet did not shed
much insight beyond this confirmation. When participants articulated specific reasons for their choice, they often did so by saying that equalizing the particularly property captured by the model they chose was more fair or less biased (34.1% of all explanations). For example, P452 (FP-Outcome-Min) said they chose the model that equalized outcomes because in the alternative model, “White people are unfairly being given bail more than blacks.” An additional 26.3% of explanations were too vague to provide additional insight. For example, P42 (Acc-RaceUsage-Maj) wrote, “A combination of both can provide valuable information. Not including race changes the results significantly. Testing both models and determining how much each factor should weigh would help improve both models.”

While few participants described fully what they meant by the terms, accuracy (mentioned in 18.9% of all answers) the use of race (13.3%), and equality (10.8%) were all invoked. Explanations invoking equality tended to further mention evenness and consistency. For example, P460 (FP-Outcome-Min) preferred to equalize outcomes because the “probability of bail is consistent across race.”

When race usage was part of the trade-off, participants’ justifications often discussed this property (24.4% of answers). For example, P245 (FP-RaceUsage-Min) preferred to equalize false positives even when explicitly considering race “because it is much more accurate overall.” Similarly, P392 (Outcome-RaceUsage-Min) preferred to equalize outcomes even when explicitly considering race because it “uses more information to make its decision and will likely have a more accurate response.” In contrast, P25 (Acc-RaceUsage-Maj) wrote, “I’d rather not have race as an issue for the computer to factor in because it is irrelevant to the decision being made.”

3.2.3 Human Judges vs. Models

After participants saw both sides of the trade-off, we asked them whether they preferred a human judge or their choice of either of the two models they had seen. For eight out of
the twelve conditions, we observed a statistically significant preference for human judges. Figure 3.3 shows this graphically. Among these eight conditions, the percentage of participants probably or definitely preferring a human ranged from 58.3% (Acc-RaceUsage-Maj) to 80.5% (FP-Acc-Maj).

We did not observe a statistically significant preference for using an ML model in any condition. The Outcome-RaceUsage conditions had the highest percent of participants who favored the model over a judge (44.4% for Outcome-RaceUsage-Maj and 41.5% for Outcome-RaceUsage-Min). Furthermore, these same conditions also had the highest proportion of people who strongly favored the model over a judge (19.4% and 19.5% for -Maj and -Min, respectively). That human judges were fallible and models were more objective appeared in 47% of participants’ free-text justifications in these cases where the participant preferred a model over a judge. For example, P372 (Outcome-RaceUsage-Maj) wrote, “Human beings are biased and easily manipulated.” Similarly, P392 (Outcome-RaceUsage-Min) wrote, “Models lack bias and personal experiences that they will not be able to use to make decisions.” However, being able to use individualized judgment was also sometimes expressed positively.
In contrast, participants expressed a number of reasons for preferring a human judge. 76.7% of justifications mentioned a judge’s ability to make individualized, case-by-case decisions. For example, P94 (FP-Acc-Maj) wrote, “I would use a human judge because I think that they would be able to see past the race of the offender and recognize outliers in their personalities and past that would make them more risky to not appear before trial.” Similarly, P127 (FP-Acc-Maj) wrote, “A human judge would be able to take race out of the equation.” An additional 29.1% of justifications discussed the judge as either more accountable or more ethical. Finally, 27.4% of justifications expressed that judges have ethical reasoning capabilities beyond those of models. According to P232 (FP-RaceUsage-Min), “I still think human judges can see things that a model cannot such as morals, values, and attitude. These cannot be taken into account by the model as it only looks at conditions known to it.” Similarly, P384 (Outcome-RaceUsage-Min) wrote, “Programmers aren’t trained to at least try not to be biased.”

Even though participants preferred equalizing false positive rates to equalizing accuracy (as described in the previous section), they nonetheless still significantly preferred a human judge. In particular 65.9% of FP-Acc-Min participants and 80.5% of FP-Acc-Maj participants preferred a human judge to either model they saw.

### 3.2.4 Fairness and Bias

Prior to having participants compare the two models in their trade-off, we asked them to rate their fairness, bias, and usefulness of each model individually. Across most models, participants tended to lean towards rating the model as “not at all biased” or only “somewhat biased.” For only one model did participants significantly lean toward rating it as biased. Nonetheless, participants tended not to rate these models as fair. In particular, for only one model did participants significantly lean toward rating it as fair. In contrast, for six others, they leaned towards rating it as “not at all fair” or only “somewhat fair.” In spite of these
trends, there were no models where participants’ opinions were unanimous, and many where opinions were strongly divided.

Bias

Participants rated the bias of each of the two models they saw on a five-point scale from “completely biased” to “not at all biased.” There were six trade-off pairs, two possible disadvantaged groups, and two models per pair, yielding twenty-four individual models. For sixteen of these twenty-four, participants significantly tended towards rating the model as not biased, as shown in Figure 3.4. For these sixteen models, the percentage of participants who rated the model as either “not at all biased” or only “somewhat biased” ranged from 56.3% to 73.9%. In contrast, for only one model (≠ FP, = Outcome, Min) were responses
Figure 3.5: Participants' ratings of each model's fairness.

significantly skewed toward rating the model as biased. That is, participants tended to consider it biased when outcomes were equalized at the expense of disparate false positive rates that disadvantaged African-American defendants.

We also tested differences between bias ratings within a trade-off pair. Consistent with the results of the preferred side of trade-offs, we found a significant difference in bias assessments between models in the FP-Acc-Min trade-off ($W = 106, p = 0.034$). That is, participants felt that equalizing accuracy was more biased than equalizing false positive rates.

Fairness

As shown in Figure 3.5, we found fewer significant trends in participants' fairness ratings. For six of the twenty-four models, participants significantly tended toward rating the model
as not fair. For these six models, between 63.6% and 87.8% of participants thought the model was “not at all fair” or only “somewhat fair.” In contrast, for only one of the twenty-four models did we observe a significant trend toward considering a model fair. Overall, 80.6% of participants rated as “mostly fair” or “completely fair” the model in which more African-American defendants than White defendants were granted bail even though race was not used as an input to the model (≠ Outcome, Race not used, Maj). This perception of fairness did not persist when White defendants were favored (see Figure 3.5). Differences based on which group was disadvantaged are discussed further in Section 3.2.5. Like bias ratings, fairness ratings were polarized. 85% of participants rated at least one of the two models they saw as either “completely fair” or “not at all fair.”

We also tested for differences in perceived fairness between models in a pair. The model equalizing false positives was rated as more fair than the model equalizing accuracy when Whites were disadvantaged (W = 289.5, p = 0.043). This echoes a similar finding in differences of their bias ratings.

The Relationship Between Bias and Fairness

While one might assume that models that are not biased are thereby fair, we observed a much more complex and nuanced relationship between participants’ ratings of bias and fairness. Participants tended to rate the model with equal outcomes and unequal false positive rates disadvantaging African-Americans (≠ FP, = Outcome, Min) as biased and not fair. Participants tended to rate the model where race was not used, but outcomes were unequal disadvantaging White defendants (≠ Outcomes, Race not used, Maj) as both fair and not biased. Participants tended to rate the model with equal false positives and unequal accuracy disadvantaging African-American defendants (= FP, ≠ Acc, Min) as not biased, yet also not fair.

We found that participants tended to rate models they considered biased as unfair.
Figure 3.6: Correlation between ratings of fairness and ratings of bias for each model. Percentages are of total answers.

Graphically, the triangular nature of Figure 3.6 shows that models that were considered biased were almost never considered to be fair. The most common combination was that a model was “not at all biased” and also “completely fair.” However, a model that they rated as not biased was not necessarily rated as fair. 39.0% of participants rated at least one model as either “not at all” or only “somewhat” biased, yet “not at all” or only “somewhat” fair. As shown in the leftmost column of Figure 3.6 some participants rated models as “not at all biased,” yet only “mostly,” “somewhat,” or “not at all” fair. That is not to say that there is not a strong association between bias and fairness. The Kendall’s $\tau$ correlation coefficient between fairness ratings and bias ratings was $-0.513$ ($p < .001$).
Ratings of a model being not biased, but also not fair, were most frequent in response to $\neq$ FP, $=$ Outcome, Maj, composing 42.1% of all responses to that model. Similarly, they were 36.6% of all responses to the $\neq$ FP, $=$ Acc, Maj model. Such answers tended to express ambivalence. For example, P447 (FP-Outcome-Maj) wrote that the model “seems somewhat fair but mistakenly denying bail to 35% of Whites and African Americans still seems like a high error rate.” They then explained that they did not think the model was biased because “Model X mistakenly denies bail equally across both Whites and African Americans.” In other words, high error rates could make a model unfair even though these rates were equal across racial groups (making the model unbiased).

The Relationship Between Bias and Judge Preferences

Initially, we had expected that participants who rated a model they saw as “not at all biased” would prefer model over a human judge since they had a more unbiased option. However, most of these participants nonetheless preferred a human judge over either model they saw. Figure 3.7 shows little difference in the distribution of preferences for a human judge versus a model between participants who rated both models they saw as unbiased and those who rated one model as biased (and the other as unbiased). About half of participants in each case “probably” or “definitely” preferred a human judge. Participants who reported both models they saw as at least somewhat biased were even more likely to prefer a human judge.

It is also noteworthy that even though 80.6% of participants thought that the $\neq$ Outcome, Race not used, Maj model was “mostly” or “completely” fair, only 44.4% of participants “probably” or “definitely” preferred a model instead of a human judge. This suggests that even relative consensus about the fairness of a model may be insufficient to produce consensus about whether an ML model is more appropriate to use than a human judge.

Qualitative coding of the reasons for bias ratings showed that participants largely understood the trade-offs, yet did not provide much deeper insight. Most frequently, participants
Figure 3.7: Preference for a human or a model, broken down by the participant’s ratings of the individual models’ bias.

reported finding a model biased because of disparate false positive rates (43.4%), the explicit consideration of race (34.7% of responses), disparate accuracy (33.5%), and disparate outcomes (29.7%).

Usefulness

Participants’ ratings of a model’s usefulness were largely redundant to those of bias and fairness. In the majority of models where we observed a statistically significant trend toward one side, the trends were toward the model not being useful. The one model where there was a statistically significant trend toward a model being useful was the already discussed ≠ Outcome-Race not used-Maj model. The results for models’ perceived usefulness are shown in Figure A.1 in the supplementary appendix.
For each of the six trade-offs, half of participants saw a model where the majority group (Whites) were disadvantaged by the disparate rate, whereas the other half saw a model where the minority group (African-Americans) were disadvantaged. We observed several instances in which there appear to be differences in participants’ reactions between the -Maj and -Min variants.

The model in which outcomes were equalized at the expense of disparate false positive rates was rated as “mostly biased” or “completely biased” by 68.9% of participants when African-Americans were disadvantaged, yet only by 39.5% of participants when Whites were disadvantaged. Similarly, in the model where race was not considered in the model and outcomes were disparate across groups, 80.6% of participants rated the model as “mostly fair” or “completely fair” when Whites were disadvantaged, yet only 51.2% did so when African-Americans were disadvantaged. One possible explanation is that when the model disadvantages African-American defendants, people are more likely to think the model is biased even when race is not explicitly used as a model input.

Despite these apparent differences, statistical testing did not distinguish between the -Maj and -Min variants. The Mann-Whitney U test comparing fairness ratings between the \( \neq \) Outcome, Race used, Maj and \( \neq \) Outcome, Race used, Min groups failed to reject the null hypothesis \( (U = 880.0, p = 0.13) \). We found a similar result for \( \neq \) FP, = Outcome Min tested against \( \neq \) FP, = Outcome, Maj \( (U = 931.0, p = 0.44) \). Note, however, that these experiments should be repeated with a larger sample size to determine whether these apparent differences are statistically significant.

Very few participants explicitly discussed which group was disadvantaged when writing free-text justifications of their choices. A post-coding review of justifications of bias ratings of \( \neq \) Outcome, Race not used showed that 16.7% of participants mentioned the group being disadvantaged. However, the discussion was mostly at the level of noting the group being
disadvantaged. As P479 (Outcome-RaceUsage-Maj) wrote, “It seems to grant Black folks bail more often and White folks less often.” We identified ten instances (out of 502 total) where the participant justified their choice of one model over the other in terms of the disadvantaged group, but there were not enough to make reliable conclusions. For example P120 (FP-Acc-Maj) wrote, “I would choose [the model with higher accuracy for Whites] since I am a white American and that model has a higher success rate for white Americans.”

A small portion of participants (4.7%) discussed the system as a whole as unjust in their explanations of why they believed a model was biased. These explanations possibly allude to racist policing and criminal justice practices. As P439 (FP-Outcome-Maj) wrote, the model they saw “disproportionately impacts while people. On the other hand, the whole bail system disproportionately impacts Black folks, so it may be a wash.”

Similar to procedural fairness, 28.7% of participants discussed in general terms whether a feature or types of data were inherently fair to use. Justifying a model that did not consider race explicitly, P61 (Acc-RaceUsage-Min) wrote, “They are using offenses and the number of the crimes to make a prediction, not by the color of the persons skin.” In a similar context, P22 (Acc-RaceUsage-Min) wrote, “I think the model just looks at the statistics and reports it how it is.” Across conditions, many participants expressed beliefs that people should be treated equally.

3.3 Discussion

After discussing our protocol’s limitations (Sec. 3.3.1), we compare our findings with those of prior empirical studies on perceptions of fairness (Sec. 3.3.2). We conclude by recapitulating our work’s key lessons and proposing future work (Sec. 3.3.3).
3.3.1 Limitations

Our experiment was limited in a number of ways. First, we studied a convenience sample that is not necessarily generalizable to any larger group. The purpose of this study was not necessarily to identify the true perceptions of the entire American public, but rather to assess whether an interest in and understandings of decisions exists outside of technical communities. Though our findings about accuracy and false positives are notable enough to bear further consideration, we expect that the precise distributions will vary if this experiment were repeated with an eye to target more specific communities.

We asked about differences in how models treated White and African-American defendants, but did not have a sufficiently large number of non-White participants to meaningfully determine whether there was an interaction between the participant’s own demographics and what group was disadvantaged. In addition, we only examined fully automated models and fully manual human judges, whereas a hybrid approach (a human who relies in part on an automated model) is a potential compromise. Furthermore, we investigated only a single visualization of the model properties and differences. While we piloted these visualizations using cognitive interviews to verify their intelligibility for participants without particular statistics expertise, other work has used different visualizations, including simple text statements of percentages, tables, and novel visualizations. We used straightforward bar graphs, annotated with the text percentage. We chose bar graphs because they communicate difference and magnitude.

3.3.2 Comparison with other empirical studies

As noted in the previous section, prior empirical studies of how humans perceive the fairness of machine learning models vary from each other, and from our work, in how they visualize the properties of these models. Our supplementary appendix discusses this confound further. While this limits our ability to directly compare with prior work, in this section we highlight
similarities and differences in conclusions among these studies. Future work could investigate how the visualization of models impacts perceptions of fairness.

Grgić-Hlača et al. found that the fairness of explicitly considering race to predict recidivism risk depended on how doing so impacted model accuracy [31]. While only 21% of their participants thought it unconditionally fair to consider race, 42% thought it fair to consider race if it improved accuracy. We investigated the fairness of explicitly considering race in a model that equalized accuracy across racial groups, rather than strictly increasing accuracy. While we observed high variance in fairness ratings for this model, most participants who rated the model as not at all fair mentioned the unfairness of using race, regardless of equalized accuracy. Among those who rated the model as fair, most referenced equalized accuracy as a mitigating factor (e.g., P12 wrote, “It’s equally accurate for people of different races, so I think that makes the use of the data justified.”). Our findings are therefore consistent with Grgić-Hlača et al.’s findings, suggesting that people sometimes find explicitly considering race justified if doing so improves performance.

Saxena et al. found the race of the person advantaged in loan allocation affected perceptions of fairness [53]. Specifically, participants found it more fair to give the entire sum to the candidate with the higher loan repayment rate than to divide it equally, but only when the candidate with the higher repayment rate was Black. In contrast, we did not observe any statistically significant effects when we compared conditions in which Whites were negatively impacted to those in which African-Americans were negatively impacted. This suggests that race may be more significant when examining individuals (as in Saxena et al.), rather than groups.

Srivastava et al. found a preference for demographic parity (equalizing the percentage of people who receive a positive classification, which was our = Outcome condition) over other definitions of fairness, like equalized false positive rates or false negative rates [55]. In contrast, we found that equalizing the false positive rate at the expense of having disparate
outcome rates across groups was preferred over the opposite. One possible explanation for this lies in the different ways the two studies visualized the properties of a model, which we discuss further in the supplementary appendix.

Kennedy et al. found the size of the training data, the false positive and false negative rates, and the institutional source most impacted which model participants trusted [41]. We also found that equalizing false positive rates was generally valued over equalizing accuracy. Whereas Kennedy et al. found that their participants generally expressed trust in algorithmic methods, our participants expressed a general preference for a human judge. A possible explanation for this difference is that while we showed differences between model properties by racial group, Kennedy et al. investigated only the overall false positive rate, false negative rate, and accuracy. Had their participants been aware of differences across racial groups, they may have been less likely to trust algorithms.

### 3.3.3 Conclusions and Future Work

Our survey-based experiment asked participants to comparatively evaluate two models that exemplified the two sides of the realistic trade-offs between fairness-related properties. We observed a marginal, yet statistically significant, preference for equalizing false positives across demographic groups over equalizing accuracy. For other trade-offs, we observed at most a weak (and non-significant) preference. Notably, though, each side of each trade-off was strongly preferred by a non-trivial fraction of participants. This result casts doubt on the possibility of achieving broad acceptance across society that the right fairness decision was made among mutually exclusive, yet seemingly desirable, statistical definitions of fairness. Furthermore, we observed a general, yet not often not overwhelming, preference for a human judge over models capturing either side of the realistic trade-offs we examined. Even when participants thought that neither side of the trade-off was biased, over half of them still preferred a human judge over the model. We also found that just because a participant felt
a model was “not at all biased” did not imply that they considered the model fair. Our findings suggest that decisions are a salient manner of characterizing models to the public.

Future work should investigate how to facilitate public involvement in decisions concerning fairness. It should examine how the method of visualizing model properties affects perceptions of fairness. It should also further examine whether perceptions differ based on demographics. As automated decision systems are deployed more widely, efforts should engage people affected by these systems, particularly those from marginalized communities.

A machine learning developer confronted with the tough types of trade-off decisions we investigated might be tempted to crowdsource the decision by surveying the public. Our findings suggest that crowdsourcing is unlikely to produce consensus or full clarity about the decision. If the developer were to then take a majoritarian approach, this process is likely to perpetuate historical discrimination and marginalization. Instead, how individual members of society perceive the fairness and acceptability of automated decision making should be considered holistically as part of the development of these systems. This study indicates that more research into tools for facilitating deliberation about models is warranted.
References


[27]


[46] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. How We Analyzed the COMPAS Recidivism Algorithm, May 2016.


A.1 Additional Figures

Figure A.1: Participants’ ratings of the usefulness of each model.
Figure A.2: Fairness ratings of models broken out by whether participant identified as white.

A.2 Demographic Models

A.3 Survey Instrument

Introduction

Thank you for choosing to participate in this study. After a short tutorial about machine learning, you will be asked to react to a scenario. After this, we will ask you some demographic questions to better understand your responses. In this scenario you will be asked to play the role of a data scientist. Specifically you will be asked to use your judgment about which model to use for a particular task.

- Machine learning models can help you use past data to make predictions.
  - **Training Data**: To make a model that predicts if a college basketball team will make the tournament, you could first gather data from past seasons about the characteristics, or features of teams that did or did not make the tournament.
Table A.1: Random effects model coefficients for model judgments. Standard error for each coefficient is reported in parentheses.

<table>
<thead>
<tr>
<th></th>
<th>Bias</th>
<th>Fairness</th>
<th>Usefulness</th>
</tr>
</thead>
<tbody>
<tr>
<td>income.L</td>
<td>-1.10 (1.31)</td>
<td>1.11 (1.17)</td>
<td>0.10 (1.18)</td>
</tr>
<tr>
<td>education.L</td>
<td>1.73 (1.01)</td>
<td>-0.76 (0.91)</td>
<td>0.29 (0.93)</td>
</tr>
<tr>
<td>raceblack</td>
<td>0.37 (0.23)</td>
<td>-0.26 (0.23)</td>
<td>-0.11 (0.23)</td>
</tr>
<tr>
<td>raceasian</td>
<td>0.05 (0.24)</td>
<td>-0.25 (0.24)</td>
<td>0.04 (0.23)</td>
</tr>
<tr>
<td>racelatino</td>
<td>-0.12 (0.29)</td>
<td>0.22 (0.30)</td>
<td>-0.40 (0.30)</td>
</tr>
<tr>
<td>racenone</td>
<td>1.37 (0.71)</td>
<td>-0.77 (0.71)</td>
<td>-0.44 (0.66)</td>
</tr>
<tr>
<td>tech_exp.L</td>
<td>-0.19 (0.24)</td>
<td>0.33 (0.24)</td>
<td>0.05 (0.22)</td>
</tr>
<tr>
<td>politicalIndependent</td>
<td>-0.21 (0.16)</td>
<td>0.11 (0.15)</td>
<td>0.01 (0.15)</td>
</tr>
<tr>
<td>politicalOther:</td>
<td>-0.30 (0.50)</td>
<td>-1.03 (0.51)*</td>
<td>-1.60 (0.57)**</td>
</tr>
<tr>
<td>politicalPrefer not to say</td>
<td>-0.96 (0.77)</td>
<td>1.35 (0.77)</td>
<td>0.07 (0.73)</td>
</tr>
<tr>
<td>politicalRepublican</td>
<td>-0.17 (0.17)</td>
<td>0.39 (0.16)*</td>
<td>0.11 (0.17)</td>
</tr>
<tr>
<td>genderFemale</td>
<td>-0.01 (0.13)</td>
<td>-0.16 (0.13)</td>
<td>0.05 (0.13)</td>
</tr>
</tbody>
</table>

Log Likelihood: -1337.95 -1325.11 -1332.19
AIC: 2761.90 2736.23 2750.39
BIC: 2969.35 2943.68 2957.83
Num. obs.: 920 920 920
Groups (trade-off): 24 24 24
Variance: trade-off: (Intercept) 0.22 0.29 0.21

***p < 0.001, **p < 0.01, *p < 0.05

- Machine learning models use patterns in the training data to make predictions.
  - For example, it might find that the coach’s number of years of experience combined with number of seniors on the team best predict whether the team will make the playoffs.
  - The collection of patterns is the model.

- A model might not make the right prediction for every team.
  - For example, a team with an experienced head coach and many seniors might also have an injured star player or might just have bad luck, and not make the playoffs.
• Once the model has been made, it is possible to test how it will perform by applying it to data that was set aside before building the model

  – For example, we could make the model using data from the 2015 and 2016 seasons, and see how well it works in the 2017 season.

**Start of New Scenarios**

The next series of questions will refer to this scenario. Metropolis city needs to decide to which people charged with non-violent offenses they can grant bail (and thus potentially release from jail) pending trial.

• *Training Data:* The city created two models by gathering data from past non-violent offense charges about the characteristics of people that did or did not show up for their trial.

• *Models:* The resulting models try to predict whether a person newly charged with a non-violent offense will or will not show up for their trial.

Based on the model Metropolis city chooses, they can decide which people to release pending trial:

• If the model predicts that a person **will** show up for their trial, they will be granted bail and potentially released.

• If the model predicts that a person **will not** show up for their trial, they will be denied bail, and remain in jail.

However, neither model they are considering is perfect. They each make mistakes in different ways.
Model X [Order of model X and model Y randomized]

Model X is one of the models Metropolis City is considering using. Below are two graphs showing properties of model X. The top graph shows the [accuracy] of model X. The bottom graph shows the [probability of being granted bail] when model X is used.

- if Acc a quality: Accuracy: the accuracy of the model is the rate at which the model makes correct predictions. A prediction is correct if the model either predicts that a defendant will show up for their trial and they would or that they will not show up for their trial and they would not have

- if Outcome a quality: Probability of bail: the probability of bail is the likelihood of a defendant being granted bail if the model is used

- if FP a quality: Mistakenly denied bail: A defendant who is mistakenly denied bail is one that the model predicts would not show up for their trial when they would have

- if Race usage a quality: Race is not one of the features used: the model makes decisions with no knowledge of the race of the subjects. Other features like type of offense and number of previous offenses are used as input to the model.

1. Do you think model X is fair? (Not at all fair, Somewhat fair, Mostly fair, Completely fair, Don’t know)

2. Why?

3. Do you think model X is biased (Not at all biased, Somewhat biased, Mostly biased, Completely biased, Don’t know)

4. Why?

5. Do you think model X is useful (Not at all useful, Somewhat useful, Mostly useful, Very useful, Don’t know)
6. Why?

7. Given a choice between model X and a human judge to make bail decisions, what would you choose? (Definitely model X, Probably model X, Unsure/can’t decide, Probably human judge, Definitely human judge)

8. Why?

Model Y

Model Y is one of the models Metropolis City is considering using. Below are two graphs showing the properties of model Y. The top graph shows the **percent probability of being granted bail** when model Y is used. The bottom graph shows what percent of defendants are **mistakenly denied bail** when model Y is used.

- Descriptions the same as in question about model X
- Questions the same as for model X

### Changed Answers?

1. Did seeing the second model change the answers you wished to give for the first model you saw? (Yes, No)

2. What aspects of your model would you wish to change and why?

Model Comparisons [Figure showing model X next to model Y]

1. Which model is more fair, model X or model Y? (Definitely model X, Probably model X, Models X and Y are equally fair, Probably model Y, Definitely model Y)

2. Why?
3. Which model is more biased, model X or model Y? (Definitely model X, Probably model X, Models X and Y are equally biased, Probably model Y, Definitely model Y)

4. Why?

5. Which model is more useful, model X or model Y? (Definitely model X, Probably model X, Models X and Y are equally useful, Probably model Y, Definitely model Y)

6. Why?

7. Given a choice between model X and model Y, which would you choose? (Definitely model X, Probably model X, Unsure/can’t decide, Probably model Y, Definitely model Y)

8. Why?

9. Given a choice between using a model (either model X or model Y) or a human judge to make the decision, what would you choose? (Definitely a model, Probably a model, Unsure/can’t decide, Probably human judge, Definitely human judge)

10. Why?

Graph Understanding

Below is a question relating to the following graph. **Do not use information you may have seen in other graphs in answering this question.**

[if accuracy a trade-off, figure showing a Group A having 79% accuracy and Group B having 53% accuracy]

Please select which statement(s) can be validly inferred from this graph.

1. More people in group A will receive bail than people in group B
2. Predictions about people in group A will be more likely to be correct than predictions about people in group B

3. People in group A will be more likely to be mistakenly denied bail than people in group B

4. People in group A will be more likely to be mistakenly granted bail than people in group B

[if FP a trade-off, figure showing a Group A having 24% false positives and Group B having 45% accuracy]

Please select which statement(s) can be validly inferred from this graph.

1. More people in group A will receive bail and not show up than people in group B

2. Predictions about people in group B will be more likely to be mistakenly denied bail than people in group A

3. People in group A will be more likely to be mistakenly granted bail than people in group B

4. People in group A will be more likely to receive bail than people in group B

[if outcome a trade-off, figure showing a Group A having 65% bail probability and Group B having 41% bail probability]

Please select which statement(s) can be validly inferred from this graph.

1. More people in group A will receive bail than people in group B

2. People in group A will be more likely to receive bail than people in group B

3. Predictions about people in group A will be more likely to be correct than predictions about people in group B

65
4. People in group B will be more likely to be mistakenly denied bail than people in group A

Cognitive Reflection Test

1. A bat and a ball cost $1.10 in total. The bat costs $1.00 more than the ball. How much does the ball cost?

2. If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets?

3. In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half the lake?

Demographic Questions

You will now be asked a series of demographic questions.

1. What is your age (18-24 years old, 25-34 years old, 35-44 years old, 45-54 years old, 55-64 years old, 65+ years old, prefer not to say)

2. What is your gender (Male, Female, Gender non-binary, Other not listed, prefer not to say)

3. Please select the set of categories that describe your racial or ethnic background. You may select multiple categories (American Indian or Alaska Native, Asian, Black or African American, Hispanic or Latino, Native Hawaiian or Pacific Islander, White, Prefer not to say)

4. What is your highest level of formal education? (Less than high school, High school graduate, some college, 2 year degree, 4 year degree, Professional degree, Doctorate,
prefer not to say)

5. Generally speaking do you think of yourself as a Republican, Democrat, Independent or something else? (Republican, Democrat, Independent, Other, Prefer not to say)

6. What is your annual household income? (Less than $20,000, $20,000 to $49,999, $50,000 to $99,999, $100,000 to $249,999, Over $250,000, Prefer not to say)

7. What is your experience with computer science? Please select as many or as few options that apply.

- I have taken a computer science course
- I have taken a course where computer science was mentioned as a topic or read about computer science topics online
- I have or had a job where computer science tasks were part of my job duties (I have written, documented, or manipulated code)
- I have never received computer science education or held a computer science job

8. What is your experience with machine learning? Please select as many or as few options that apply.

- I have taken a course in machine learning
- I have taken a course where machine learning was mentioned as a topic or read about machine learning topics online
- I have or had a job where machine learning tasks were part of my job duties (model training, model debugging etc)
- I have never received machine learning education or held a job in which it was used
9. What is your experience with probability? Please select as many or as few options that apply.

- I have taken a course in probability
- I have taken a course where probability was mentioned as a topic or read about probability topics online
- I have or had a job where probability related tasks were part of my job duties
- I have never received probability education or held a job in which it was used
A.4 Comparison of model visualizations with other empirical work

In work by Srivastava et al. participants were asked to make twenty comparisons generated through an adaptive algorithm meant to converge upon each participant’s preferred notion of fairness. First, we note that the examples generated under this methodology do not necessarily encode any notion of trade-off between properties. Second, the way in which the information was displayed may play a role in their finding of a preference for demographic parity. The study used a display showing stylized pictures of people with different combinations of races and genders along with the true label. Below this were two rows each containing a color coded prediction for each person (see Figure A.3). We think that this display requires a greater amount of cognitive load to compare quantities like accuracy or false positive rates than demographic parity. In order to calculate such a quantity, the survey taker would need to count and remember the number of misclassifications by group. Then the survey taker would need to calculate the rate between groups. Demographic parity is much more visually clear. A survey taker could count the number of positive classifications for each group, and since there were ten people displayed, an approximate count would suffice to move towards demographic parity. Srivastava et al. also performed a survey where they asked survey respondents to choose between three models with differing overall and intergroup accuracy. However in this experiment, they did not test qualities against one another.

By contrast, our way of depicting differences between models encodes far less information (see Figure A.4). At the same time, it is much more clear about the differences we are trying to test.

At time of submission, depictions of models used by Kennedy et al. [41] were unavailable.
Figure A.3: Visualization used by Srivastava et al. [55] in their investigation of fairness considerations.

Figure A.4: A visualization of two models from our study. This one depicts the two models in the Acc-Outcome-Min condition.