The Role of Accuracy in Algorithmic Process Fairness Across Multiple Domains

by

Michele Albach

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

Department of Computing Science

University of Alberta

Abstract

Machine learning is often used to aid in human decision-making, sometimes for life-altering decisions like when determining whether or not to grant bail to a defendant or a loan to an applicant. Because of their importance, it is critical to ensure that the processes used to reach these decisions are considered fair. A common approach is to enforce some fairness constraint over the outcomes of a decision maker, but there is no single, generally-accepted definition of fairness. Additionally, most of the literature on algorithmic fairness focuses only on one or two domains, whereas machine learning techniques are used in an increasing number of distinct decision-making contexts with differing pertinent features.

In this work, we consider six different decision-making domains: bail, child protective services, hospital resources, insurance rates, loans, and unemployment aid. We focus on the fairness of the process directly, rather than the outcomes. We also take a descriptive approach, using survey data to elicit the factors that lead a decision-making process to be perceived as fair. Specifically, we ask 2157 Amazon Mechanical Turk workers to rate the features used for algorithmic decision-making in one of the six domains as either fair or unfair, as well as to rate how much they agree or disagree with the assignments of eight previously (and one newly) proposed properties to the features. We show that, in every domain, disagreements in fairness judgements can be largely explained by the assignments of properties to features. We also show that fairness judgements can be well predicted across domains by training the

predictor using the property assignments from one domain's data and predicting in another. These findings imply that the properties act as moral determinants for fairness judgements, and that respondents reason similarly about the implications of the properties in all the decision-making domains that we consider. Although our results are mostly consistent across domains, we find some important differences within specific demographic groups in the hospital and insurance domains. However, a single property usually holds the majority of the predictive power. With some exceptions, predictors learning from only the "increases accuracy" property perform better (in all domains) than predictors learning from any combination of the other seven properties, implying that the primary factor affecting respondents' perceptions of the fairness of using a feature for prediction is whether or not a feature increases the accuracy of the decision being made.

Preface

The research conducted for this thesis was a collaboration with my supervisor, Dr. James Wright. The experimental designs described in Chapter 3 and the analyses described in Chapter 4 were collaboratively created, but carried out by myself. The review of related literature in Chapter 2 is my own work.

No part of this thesis has been previously published, however we have submitted a paper based on this work to The Twenty-Second ACM Conference on Economics and Computation (EC'21).

The surveys conducted as part of this research received ethical approval from the University of Alberta Research Ethics Board, study title "Perceptions of Fairness across Contexts," ID No. Pro00095367, approved November 18th, 2019.

To my family and to my friends,
thank you for supporting me through the ups and the downs of my educational pursuits.

...

Especially to my parents, Curtis and Catherine,
you have not only supported me, but also continuously pushed me to be my best self throughout all of my life. Thank you.

...

Finally, to my Pépère,
I miss you very much.

Acknowledgements

I would first like to thank my supervisor, Dr. James Wright, for not only guiding me early in my degree to discover areas of research that interested me, but also for teaching me your research tricks and tips, pushing me to further understand research concepts, and always prompting insightful discussions. I would also like to thank the members of the ABGT reading group for lively discussions and offering their interesting perspectives on issues relating to algorithmic fairness. I am also thankful to the professors who taught me and nurtured my interest in computing science throughout my undergraduate degree, including Dr. Denilson Barbosa and Dr. Richard Sutton.

I would like to acknowledge generous funding from the University of Alberta Faculty of Graduate Studies and Research as a Graduate Fellowship in the fall of 2019.

Lastly, I would like to thank my family and friends, especially my parents Catherine and Curtis, brother William, and partner Jeff. Thank you for supporting me emotionally and mentally, especially during a global pandemic, as well as for reading everything that I write to check for mistakes and give feedback, no matter how long or technical it is.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

As machine learning becomes increasingly prevalent in everyday life, it is often used to aid, or even largely replace, human decision-making. These decisions can have life-altering consequences, such as determining unemployed citizens' eligibility for labour market programs [68], whether to provide additional medical care [5, 10], or whether to grant bail [24]. As a result, it is critical to have confidence that these decisions are made fairly; thus, there is a growing interest in analyzing and improving the fairness of algorithmic decision-making practices.

Most researchers take a normative approach, mathematically defining fairness and enforcing these definitions over *outcomes* [e.g., 9, 15, 17, 27, 43, 88]. This approach focuses mainly on limiting disparate impacts between certain protected groups of people that may be caused by the outcomes chosen by algorithmic decision makers (ADMs), or on ensuring that good outcomes are not overly concentrated. Alternatively, some researchers have focused on ensuring that the *processes* involved in algorithmic decision-making are fair [e.g., 38, 40, 41, 46, 58]. This is often achieved by limiting or restricting the information used by ADMs.

Much of the literature in fair machine learning focuses solely or mainly on the domain of bail decisions [e.g., 9, 15, 17, 38, 40, 41, 46, 79]. This is largely inspired by Propublica's influential work [2] on a program called COMPAS [24], which aids American judges when deciding whether or not to grant bail to a defendant by predicting the chance that the person will recidivate. However,

bail is by no means the only domain in which machine learning is used to make high-stakes decisions.

In this work, we take a descriptive, process-based approach, using empirical data on the perceived fairness of using different kinds of information to reach decisions. We analyze multiple decision-making domains in order to identify and explain heterogeneity in these judgements, and attempt to uncover the principles by which those judgements are made.

Our work builds on that of Grgić-Hlača *et al.* [38] (henceforth GH18), who analyze survey responses to understand the moral reasonings used by laypeople when determining what information is fair or unfair to use in decision-making. They find a high degree of disagreement in fairness judgements, but are able to well predict how fair a respondent found a given feature using the respondent's assignment to the feature of eight properties: *relevance, causes outcome, reliability, causes disparity in outcomes, caused by sensitive group membership, causes vicious cycle, volitionality,* and *privacy.* Their results imply that, although there is much disagreement on what information is fair or unfair to use as a feature, if consensus could be reached on what properties are held by each feature then people might agree about the feature's overall fairness as well. This work is an important step towards ensuring process-based fairness in ADMs, but since the analysis covers only one domain it is uncertain if the results can be generalized to other decision-making contexts.

In this paper, we begin by analyzing six decision-making scenarios to explore whether or not moral reasonings differ across domains. The domains that we use are *bail, child protective services, hospital resources, insurance rates, loans,* and *unemployment aid.* Survey respondents were each assigned to a single domain and asked how fair it is to use features associated with that domain when making decisions. Additionally, they were also asked to rate how much they agreed or disagreed with statements assigning each of the eight properties suggested by GH18 to the features. In all six domains, we find disagreement in both the fairness judgements of the features and the assignments of properties to features, although, in line with the work of GH18, we are able to accurately predict the differences in fairness judgements using

the property assignments. Between domains, we find that remarkably similar relative importances are placed on the properties, and furthermore we are able to predict fairness judgements by training on the property assignments from completely separate domain data, achieving accuracies around 80%.

However, in most cases, the single property *relevance* is able to predict fairness judgements of features better by itself than all other seven properties together.[1] In fact, *relevance* alone frequently predicts as well as or better than all eight properties together.[2] We investigated this result further with two additional survey studies. In one, we replaced questions about the property *relevance* with questions about the more specific property *increases accuracy*; in another we simply eliminated the property *relevance* entirely. Taken together, our results from these surveys support the conclusion that the most important property used by survey participants to reason about the fairness of using a feature is whether or not a feature increases prediction accuracy.

While analyzing all of our data, our conclusions are consistent across all six domains. However, after removing privileged demographic groups, we find important differences relating to the predictive powers of the properties in the hospital and insurance domains. After removing Caucasians from our data in the hospital domain, the properties *relevance* and *increases accuracy* are no longer better at predicting perceived fairness than the other properties. Instead, the property *causes outcome* predicts fairness judgements of features better by itself than all other seven properties together, but only using data from the survey that replaced *relevance* with *increases accuracy*. After removing men from our data in the insurance domain, the property *increases accuracy* no longer predicts perceived fairness as well as *relevance* does in the initial survey. In fact, no property other than *relevance* emerges as especially important to moral reasonings about fairness judgements by our female participants in the insurance domain. Future work is needed to further understand the differences in these domains that are introduced by participants'

[1] This is true in both our own data, and in the data collected by GH18.

[2] The worsened accuracy when additional properties are added is possibly due to overfitting.

demographic features.

After thoroughly surveying background materials and related work in the next chapter, we begin by describing our methodology in detail in Chapter 3, including the six domains we consider, the potential features within those domains that participants were asked about, and the eight properties used to predict fairness judgements about a feature. We performed two pilot studies: one of our own design to check whether participants were overly framed by the focus on "machine learning" rather than generic decision-making procedures, and another replicating GH18's check that the properties are both necessary and sufficient to fully explain participants' reasoning. Next in Chapter 4, we begin by quantifying the degree to which respondents' disagree about both fairness judgements of the features and assignments of properties to the features. We then construct regression models for predicting fairness judgements based only upon respondents' property assignments both within and between domains. We find that these judgements can be well predicted by the single property *relevance*. Section 4.2 describes the results from our two additional studies which replace and remove *relevance* as a property and find a new single property *increases accuracy* to also be highly predictive of fairness judgements. Then, Chapter 5 describes our demographic analyses, where we find some differences in results in the hospital and insurance domains after removing privileged groups from our data. Finally, we discuss the implications of our results in Chapter 6.

Chapter 2

Background Material

As ADMs become increasingly prevalent in many people's lives, research in the field of fair machine learning has also surged in popularity. This is especially true in the last five years, and can be attested partially to a particular article published by the journalism company ProPublica in 2016 that brought light to potential issues of bias and unfairness in court-involved computer decisions [2]. The article claims that a program widely used in the American court system called the *Correctional Offender Management Profiling for Alternative Sanctions* (COMPAS), which predicts a defendant's likelihood of committing another crime if released, unfairly judges Black defendants harsher than white defendants. Propublica's article was disputed in an article published by Northpointe [24], the company behind COMPAS, which shows by their own calculations that COMPAS is not biased. Both sides treat the issue straightforwardly and claim their math to be proof of bias or unbiasedness, but in actuality the issues of fairness and discrimination caused by ADMs are much more grey and complex due to conflicting definitions of fairness and bias. Specifically, the ProPublica article claims that COMPAS is biased by showing that Black and white defendants are subject to different error rates (as in white defendants are more likely to be misclassified as low risk to recidivate and Black defendants are more likely to be misclassified as high risk), and the Northpointe response claims that their program is not biased by showing that Black and white defendants are predicted to recidivate with equal accuracy. Each side of the argument uses a different mathematical definition of bias in machine

learning. Worse, Chouldechova [15] shows that, when the true proportion of recidivating defendants (the base rate) differs across groups of defendants, the two fairness definitions used by Propublica and Northpointe cannot be simultaneously made true. So which definition should be taken as correct? Should the same definition be used in other decision-making contexts? Researchers in the field of fair machine learning have worked to answer these questions. The following chapter will survey the literature over various subcategories of fairness in ADMs. The first section will outline the many normative fairness definitions that have been proposed, their compatibility with each other, and their potential issues. The next section will move into outlining work that has descriptively compared and tested fairness definitions by asking real people for their opinions and attempting to understand what factors affect those opinions. Lastly, we will highlight research that has discussed the implications of public perceptions of fairness in machine learning, especially pertaining to levels of trust in algorithmic versus human decision-making.

2.1 Normative Proposed Fairness Definitions

In this section we will describe the many definitions and criteria for fairness in machine learning that have been proposed. These definitions are all normative, meaning they are proposed in the context of how fairness *should* be understood and enforced, as opposed to how fairness actually *is* understood by people (which will be discussed in the next section). Many of these definitions have been proposed and interpreted more than once by different researchers, so some of them have more than one name or differing definitions under the same name.

2.1.1 Calibration

A desirable, and arguably necessary, property for a predictor is that its outcomes accurately reflect true likelihoods. An ADM is said to be *well-calibrated* if: out of the people assigned a certain probability chance of having an outcome, that same proportion actually have the outcome (for example, out of

defendants given a risk score of 7/10, 70% of them actually reoffend) [56].

Across all cases, this criterion is a measure of functionality, but calibration can be used as a measure of fairness by comparing between protected (or any) groups of people. Kleinberg *et al.* [56] define *calibration within groups* to be satisfied if, between the members of two groups given the same probability of an outcome, the actual proportion of the outcome is also equal. Equivalently, Corbett-Davies and Goel [17] and Chouldechova [15] define calibration as a satisfied fairness criterion if the true outcomes are independent of protected attributes conditional on a predictor's output scores. Put simply, these definitions require that an ADM be equally well-calibrated across (protected) groups. In the debate about COMPAS, Flores *et al.* [32] show that the program is equally well-calibrated across Black and white defendants. Dieterich *et al.* [24], on behalf of Northpointe, also show that COMPAS predicts recidivism in Black and white defendants equally well using two of their own accuracy-based fairness definitions that are similar to calibration: *accuracy equity*, and *predictive parity*. Accuracy equity requires accuracy levels measured by the area under the ROC curve to be equal across protected groups.[1] Predictive parity is very similar to calibration within groups, except rather than conditioning on an output score, the scores are divided by a threshold into positive and negative predictions and required to have equal positive prediction values across protected groups[2] [15, 69]. Dieterich *et al.* [24] show that COMPAS, which outputs a risk score from one to ten, maintains predictive parity when dividing scores into positive and negative predictions using a score of larger than four or any higher threshold.

A computer predictor that is not well-calibrated between groups is arguably not a very good predictor at all. However, calibration between groups alone can be insufficient to guarantee fairness [17]. This is most easily demonstrated by the illegal practice of redlining: limiting financial services based only on

[1]The area under the Receiver Operating Characteristics curve is a commonly used measure of performance in machine learning, it has also been used as a fairness metric by Skeem and Lowenkamp [77].

[2]Predictive parity and calibration within groups are equivalent in the case of a binary output score.

geographical location as a means to deny loans to poorer and minority people who live in specific neighbourhoods (the practice is called redlining because banks could draw a red line on a map around poorer neighbourhoods with minority residents that were automatically rejected, and was made illegal in 1977 in the United States [63]). Although it may be true that residents of a specific poorer neighbourhood have a higher risk of loan default on average, it is unfair to some residents to label them as high risk using only their address when the inclusion of their income or credit history may have shown otherwise. If an ADM labelled applicants as low or high risk to be unable to pay back a loan using only their address, the predictor could still be well-calibrated across protected groups while evidently behaving unfairly. This issue can be generalized to any scenario when a predictor groups people around an average while ignoring other predictive factors. Furthermore, a form of this problem extends to any predictor that bins the assigned probabilities into discrete scores or categories. If loan applicants or defendants are aggregated into categories of low, medium, or high risk, then a predictor could be considered well-calibrated at that three-point scale even if it is not well-calibrated on a larger or continuous scale [17]. Moreover, a predictor that is well-calibrated at a particular scale may not maintain predictive parity at every possible threshold for dividing positive and negative predictions [24].

2.1.2 Classification Parity

A multitude of proposed fairness definitions are calculated using a two-by-two confusion matrix where cases are separated by their predicted and actual outcomes into four categories: true positives, false positives, true negatives, and false negatives (Table 2.1). This wide array of definitions are all grouped together by Corbett-Davies and Goel [17] under the umbrella term *classification parity*, which they define as any fairness criterion that constrains some measure of classification error to be equal across protected groups.[3]

Most notably, Angwin *et al.* [2] use a form of classification parity in their

[3]Corbett-Davies and Goel [17] also include accuracy equity and predictive parity in this category, but we consider them more similar to calibration.

	Predicted True	Predicted False
Actually True	True Positives	False Negatives
Actually False	False Positives	True Negatives

Table 2.1: A confusion matrix dividing cases into true positives, false positives, true negatives, and false negatives depending on their predicted and actual outcomes.

ProPublica article against COMPAS: they find that the program produces higher false positive rates and lower false negative rates for Black defendants than for white defendants (using a threshold score of four to divide into positive and negative predictions). The comparison of misclassification rates between protected groups has been commonly proposed as a fairness measure under many different names. According to Chouldechova [15], Angwin *et al.* [2]'s findings about COMPAS are a violation of *error rate balance* which requires false positive and false negative error rates to be equal across protected groups (Berk *et al.* [9] refer to this same definition as unequal *conditional procedure errors*). According to Corbett-Davies *et al.* [19], they are a violation of *predictive equality* which requires only false positive error rates to be equal. According to Zafar *et al.* [88], they are a violation of *disparate mistreatment* which is a broad term that is violated any time misclassification rates differ across groups. Another proposed fairness definition that relies on error rates is Hardt *et al.* [43]'s *equalized odds*, which constrains that predicted outcomes be independent of protected groups conditional on the true outcome. In other words, equalized odds requires the proportion of true positives over all actual positive outcomes and the proportion of false positives over all actual negative outcomes be equal across protected groups. Hardt *et al.* [43] also proposed a slightly relaxed fairness criterion, *equalized opportunity*, which only holds the true positive proportions to be equal (because the positive outcome is usually the more advantageous one). Lastly, Kleinberg *et al.* [56] also define fairness criteria relating to false positive and false negative rates called *balance for the positive/negative class*, although their definitions differ slightly from those listed above because they do not require a threshold to divide scores into positive and negative predictions.

Another common category of confusion-matrix-based definitions for fairness are definitions that constrain the proportion of positive predictions to be equal across protected groups. The most common term for this tactic is *statistical parity* [9, 15, 19, 27, 56], which Dwork *et al.* [27] define to be true when the demographic proportions of the populations that receive a positive and negative prediction are identical to the demographic proportions of the population as a whole. This fairness constraint has also gone by the names of *demographic parity* [17, 34, 35, 69], *disparate impact* [31, 34, 35, 66, 69, 88], *benchmarking* [76], and *equal acceptance rates* [90]. The term disparate impact is particularly prone to confusion because of its origin in United States labour law, which defines it as any action (intentional or not) that disproportionately affects members of protected groups without cause [42]. Some authors use disparate impact as a generic term for unfairness in machine learning [7, 15, 77], but Friedler *et al.* [35] and Pessach and Shmueli [69] mathematically distinguish between disparate impact and demographic parity as comparisons of the ratio[4] and difference (respectively) between proportions of positive predictions across protected groups. Additionally, Corbett-Davies *et al.* [19] define *conditional statistical parity* as an extension of statistical parity that allows for the control of some relevant risk factors prior to constraining proportions of predictions.

The appropriateness of statistical parity as a fairness metric is heavily context-dependent, because its enforcement may decrease accuracy and cause harm in cases where base rates (the true proportions of a population that exhibit an outcome) differ across groups. As an extreme example, Black infants are significantly more likely to be born with sickle cell disease than white infants [14], so it would be inaccurate and harmful to restrict a predictor to output equal proportions of positive predictions across race for sickle cell disease in infants. However, sometimes when base rates differ across protected groups it is because of historical inequalities that have previously limited dis-

[4]Often, disparate impact refers specifically to the 80% rule, a legal definition where the ratios of positive outcomes between a disadvantaged and an advantaged group must be less that 0.8 [30].

advantaged groups, meaning that completely accurate predictors will cause harm by replicating those inequalities [6, 7]. For example, it is likely that the true proportion of men who are currently skilled enough to perform well in a position as a software engineer is higher than the true proportion of women, but this is only because of societal gender norms that have pushed young girls away from pursuing technical careers. In this case, the enforcement of statistical parity may help to counteract historical inequalities despite decreasing accuracy. This tactic, called *affirmative action*[5] [27, 67], is commonly used (and sometimes enforced by law [28]) in talent-based decision contexts like hiring or university admissions. Much like calibration, statistical parity alone can be insufficient to prevent unfairness. Dwork *et al.* [27] explain, through three similar examples, how statistical parity can be maintained while unfairly making accurate predictions in one protected group and inaccurate predictions in another. This could be caused inadvertently by the use of factors that predict an outcome well in one group but not in another (for example, SAT scores could be a better predictor of academic performance in white students with more access to tutoring than in Black students [54]), or this could be accomplished maliciously for the purpose of justifying future discrimination (for example, if a racist employer purposefully invited qualified white candidates and unqualified Black candidates to an interview).

2.1.3 Individual Fairness

All of the fairness definitions described so far have been examples of *group fairness* [27] (or *statistical fairness* [51]) because they are concerned with comparing differences in ADM impact on groups of individuals, usually divided by protected attributes. Another way of thinking about fairness is to instead consider the people about which decisions are being made on an individual level. In a comparatively early paper for the field of algorithmic fairness, Dwork *et al.* [27] propose *individual fairness*, which at a high-level is

[5]The term affirmative action was coined by African American attorney Hobart Taylor Jr. and was used in the United States Executive Order 10925, which set up the Committee on Equal Employment Opportunity [65].

maintained when similar individuals (relative to the decision being made) are treated similarly. Dwork *et al.* [27]'s method assumes the prior existence of: a similarity/distance metric, which takes as input any two individuals (or any two outcomes) and outputs the distance between them; and a loss function, which is defined to encapsulate the goals and purpose of the decision maker. Then, the problem of finding a fair predictor is framed as minimizing the loss function while constraining that the distance between any two outcomes is less than the distance between those individuals. Additionally, Dwork *et al.* [27]'s method can optionally constrain statistical parity to be maintained for the purposes of affirmative action. Other researchers have put their own spin on individual-based fairness criteria. Joseph *et al.* [48] (and companion paper [49]) define *Rawlsian fairness*, based on Rawls [73]'s theory of justice, so that an individual's probability of having a desired outcome (for example: being granted a loan) is always higher than the probability that any given less-qualified individual receives the outcome. Kusner *et al.* [58] propose a form of individual fairness that they call *counterfactual fairness*, which guarantees that individuals are treated identically to how they would have been treated in a counterfactual world where they belonged to a different demographic group.[6] Finally, Speicher *et al.* [78] expand individual fairness from a binary constraint to a measure of unfairness using the economic concept of inequality indices. Individual fairness definitions are an interesting alternative to group fairness definitions, however many of them may be harder to implement in practice due to their strong assumptions of available metrics.

2.1.4 Intersectional Fairness

As a means of bridging the gap between group fairness and individual fairness, Kearns *et al.* [51] propose that group fairness constraints should be enforced not only between high-level protected groups like race or gender, but also between intersecting subgroups of protected features. They describe how, without checking for *subgroup fairness*, decision makers are not prevented

[6]Counterfactual frameworks have been applied to other fairness definitions as well by Altman *et al.* [1], Khademi *et al.* [53], and Qureshi *et al.* [72].

from *gerrymandering*: an intentional or unintentional result where only certain members of each high-level group are given a positive outcome. Foulds *et al.* [33] point out that this issue is closely related to the feminist concept *intersectionality*, introduced by Crenshaw [20], which she demonstrates through the example court case DeGraffenreid v General Motors [22]. In 1976, five Black women in the United States sued the company General-Motors on the basis of the 1964 Civil Rights Act because the company had not been hiring any (or only hiring and then laying off) Black women. Their case was rejected because the court found that the company was not guilty of: sex discrimination, because they had hired white women; or race discrimination, because they had hired Black men. The court refused to consider that the intersecting discriminatory effects of being Black and female were greater than the sum of the sexist and racist effects alone. To prevent gerrymandering like this case in ADMs, researchers have adapted group fairness definitions to be enforced within protected subgroups (like Black women). Kearns *et al.* [51] define *statistical parity subgroup fairness* and *false positive subgroup fairness*, which respectively adapt statistical parity and false positive error rates; Hébert-Johnson *et al.* [45] define *multicalibration* which is an adaptation of calibration; and Foulds *et al.* [33] define *differential fairness* which adapts disparate impact to be enforced within subgroups.

2.1.5 Process Fairness and Anti-Classification

Most often, fairness definitions proposed by machine learning researchers are concerned with the distribution of *outcomes* received by groups or individuals. According to justice theories [57], these *outcome fairness* criteria are examples of *distributive justice*. Another important part of justice theory that common fairness criteria mostly ignore is *procedural justice*, which is concerned with the *processes* involved in decision-making as opposed to the outcomes. Grgić-Hlača *et al.* [40] introduced the term *process fairness* to describe process-based definitions of algorithmic fairness. Despite receiving less attention in research communities, a form of process fairness is arguably the most commonly suggested solution: removing protected attributes from data. For example, many

proponents of process fairness would advocate for features like race, gender, and age to be inaccessible to ADMs, a technique that Corbett-Davies and Goel [17] refer to as *anti-classification*. However, the simple removal of protected attributes is not enough to prevent discrimination because other relevant features may act as dependent proxies for the protected attributes, allowing ADMs to accurately guess them anyways. Altman *et al.* [1] show using counterfactuals that the removal of race as a predictive feature in bail decisions does not significantly alter a defendant's life course after sentencing. To account for this problem, some researchers have suggested pre-processing methods that remove all proxies and create a dataset that is independent of protected attributes [e.g., 13, 31, 46, 47, 89]. A major problem with this method however, which we will discuss further in the next section, is that as more features are removed from data, overall levels of prediction accuracy go down as well [13, 40, 41].

2.1.6 Impossibilities and Trade-offs

If all of the above mentioned fairness criteria could be maintained simultaneously, without loss of accuracy, then the field of fairness in machine learning would be much less controversial. After the initial awareness-drawing dispute between Propublica and Northpointe, multiple researchers responded with proofs of impossibility theorems between fairness definitions [15, 18, 56, 71]. Chouldechova [15] proves most succinctly that the exact fairness definitions used by the two companies, error rate balance and predictive parity, are incompatible when the true base rates differ across groups (which is true in the COMPAS data set). Kleinberg *et al.* [56] and Pleiss *et al.* [71] also each prove a similar impossibility theorem between error rate balance and calibration when base rates differ. More specifically, they show that error rate balance and calibration can only be simultaneously achieved in the cases of perfect prediction accuracy (which we do not have or this would not be a topic of interest) or equal base rates. Worse, Kleinberg *et al.* [56] also show that the two definitions cannot even be approximately achieved unless base rates are approximately equal. Pleiss *et al.* [71] do find that calibration can be main-

tained while holding a single error rate to be equal (for example, Hardt *et al.* [43]'s equalized opportunity), but that the other error rate will necessarily be made worse. Because of the controversy surrounding COMPAS, much work has focused specifically on the definitions of calibration and error rates, but incompatibilities exist between other fairness definitions as well. Narayanan [67] claims that Chouldechova [15]'s proof could be completed with any three classification parity definitions (that can be computed using a confusion matrix). Corbett-Davies and Goel [17] show that any decision maker that uses a single threshold policy to allocate outcomes (which is related to individual fairness and the concept of similar individuals being treated similarly) will violate classification parity definitions.

The aforementioned impossibility theorems have something in common: they are all between outcome-based definitions. In fact, there is a general root problem with using outcome-based bias testing in decision-making when base rates differ, called the problem of *infra-marginality* [4, 17, 76]. The problem of infra-marginality concerns the disconnect between outcome tests that evaluate statistics that are away from a threshold of decision-making (in other words, infra-marginal statistics), as opposed to societal and legal unfairness concerns that usually evaluate decisions made close to a threshold. Specifically, whenever a single threshold is used for two groups with differing base rates, infra-marginal statistical values will also differ (Figure 2.1). So, outcome tests that find discrepancies in statistical values between groups might only be evidence of differences in base rates, not evidence of disparity or unfairness[7] [17]. As a solution to this problem, Simoiu *et al.* [76] propose the *threshold test* to replace outcome tests. The threshold test uses Bayesian inference to determine the actual group-specific thresholds used by decision makers, and then, drawing from individual fairness, concludes that discrimination occured if and only if the thresholds differ between groups. Additionally, Žliobaitė [90] draws attention to a similar problem relating to the use of statistical parity as an outcome test when base rates differ, and offers her own solution by

[7]Although whether or not differences in base rates are themselves signs of disparity is a different debate, one which is argued for by Foulds *et al.* [33].

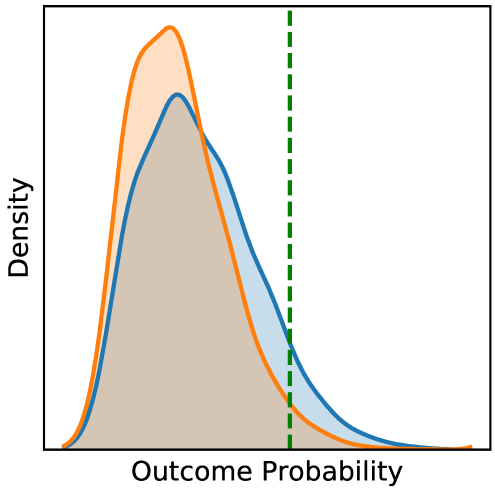

Figure 2.1: A visualization of the problem of infra-marginality. Two hypothetical true risk distributions over some predicted outcome are different (representative of how true base rates often differ between groups in real data). When the same threshold (green line) is applied to both of them, then levels of precision, represented by the size of the tails to the right of the threshold, will also differ. This problem persists for other infra-marginal statistics as well, like false positive rates. Figure adapted from Corbett-Davies and Goel [17].

normalizing the discrimination measure with respect to the acceptance rate.

Another hugely important trade-off to be considered by ADM developers, one which is highly referenced in the fairness literature, is the supposed *accuracy-fairness trade-off* [e.g., 8, 9, 13, 40, 41, 55, 66, 90]. Accuracy levels can be diminished by the enforcement of both outcome and process fairness definitions. Relating to fair outcomes, this trade-off is closely tied to the problem of infra-marginality because it is again caused by differing base rates. When true base rates are different between groups, then enforced fairness criteria like statistical parity or error rate balance will lower overall accuracy levels [8, 9]. However, the reversal of this trade-off presents a different perspective to be considered: when true base rates differ, then a perfectly accurate predictor will be labelled unfair by some fairness criteria. This perspective brings back the context-based question of why base rates differ in the first place, and relates to discussions of affirmative action from earlier sections. Relating to

fair processes, the accuracy-fairness trade-off presents itself through the closely related trade-off between accuracy and simplicity. It is often considered that an ADM is more fair if it is easier to understand how the decision is being made. As a method, process fairness (specifically anti-classification) is closely tied to algorithm interpretability because it involves removing features, and an algorithm is more simple to understand the less features that it has. However on the other hand, an algorithm can usually be more accurate the more features that it has, resulting in a trade-off. Multiple researchers have found that implementing measures of process fairness decreases accuracy [13, 40, 41], but they argue that it is a necessary sacrifice to improve fairness. Conversely, Kleinberg and Mullainathan [55] develop a model that shows that simplicity in ADMs is not tied to fairness, but actually lowers overall welfare levels for disadvantaged groups.

Considering all of the impossibility theorems and trade-offs presented above, developers must make difficult choices about what definitions to follow when implementing ADMs. Researchers have varying suggestions of what to prioritize: some suggest taking into account the differing costs of fairness constraints [56], while others suggest not striving for a single correct definition at all [67]. Relating specifically to the accuracy-fairness trade-off, some researchers claim that as computer programmers our job should be to create predictors as accurate as possible and leave fairness concerns to policymakers [8, 9, 17], while others have questioned whether or not achieving high accuracy should be a decision-making goal at all when predictors are trained on data resulting from historical discrimination [31, 33]. Overall, the vast number of proposed fairness criteria makes it important for professionals of varying expertise to come together and consider the implications of what fairness definitions to use when algorithmic decision-making affects real people's lives.

2.2 Descriptive Analyses of Fairness Definitions

So far, the fairness definitions discussed have been normatively proposed, suggesting how ADMs *should* make decisions. Another important direction for research is to descriptively study whether or not regular people actually *consider* these definitions to be fair. Especially considering the multitudes of trade-offs and impossibilities that exist between definitions, a natural next step is to determine what real people (especially the people whom these decisions effect) prioritize when it comes to fairness. In this section, we will begin by outlining descriptive survey work that has provided support for normative definitions, whether by directly asking participants to choose between contrasting definitions, or by using participant answers as tools to enforce fairness. Then, we will cover research that has worked to understand how people perceive fairness, as well as the effects that those perceptions of fairness can have on issues of importance to implementers of ADMs. Lastly, we will discuss differences in perceptions of fairness (and trust) when decisions are computer-made as opposed to human-made.

2.2.1 Choosing Between Fairness Definitions

Some researchers have taken the simple approach of directly asking survey respondents to choose between fairness definitions in specific scenarios, although results are often mixed or contrasting. Saxena *et al.* [74] do so in the context of loan decisions by asking survey respondents to rate the fairness of possible decision outcomes. The outcomes conform to three different fairness definitions: individual fairness [27]; Rawlsian fairness [48]; and calibration [60], and the results show a preference for calibration. Srivastava *et al.* [79] compare a more diverse set of fairness definitions in the context of recidivism prediction. Using participant's labellings of specific outcomes as discriminatory or not, Srivastava *et al.* [79]'s adaptive experiment pinpoints each participant's most compatible fairness definition, which turns out to be demographic parity for most participants (even when more complex definitions are thoroughly

explained). Lastly, Harrison *et al.* [44] asked survey respondents to choose between two hypothetical models, also in the context of recidivism prediction, where one fairness definition was satisfied and another violated. Their participants were assigned two of the four fairness definitions: equalized accuracy; equalized false positive rates; equalized outcomes; and the consideration of race, and asked to choose between decision rules that prioritized one over the other. Harrison *et al.* [44] find some preference for equalized false positive rates over equalized accuracy, but find mostly inconclusive results since non-trivial amounts of respondents preferred each side of every trade-off. Overall, researcher's attempts to determine descriptive preferences between fairness definitions have not resulted in any consistent consensus of a ground truth, but this could be due to the varying contexts and definitions that have been compared.

2.2.2 Descriptively Enforced Anti-Classification

Anti-classification, the process-based fairness definition of not allowing certain features to be used by decision makers if they are deemed unfair, is normative as a general concept, but becomes descriptively defined when real people's opinions are used to determine which features are fair or unfair to use. Grgić-Hlača *et al.* [40] (and Grgić-Hlača *et al.* [41]) explore this concept by defining three forms of feature-specific process fairness that each depend on participants' opinions of features given different outcome effects. First, they define *feature-apriori* fairness as the proportion of people who consider a feature to be fair to use by an ADM prior to learning anything about the feature's effects on outcomes. Then, they define *feature-accuracy* and *feature-disparity* fairness to be the new proportions of people who consider a feature fair if using that feature increases the accuracy or the disparity (respectively) of a classifier's outcomes. Grgić-Hlača *et al.* [40] apply their three feature-specific fairness definitions to the context of recidivism prediction using nine hand-picked features from the COMPAS dataset. They surveyed Amazon Mechanical Turk (AMT) workers, asking first (for each of the nine features) whether participants believed that it is fair or unfair to use a feature when estimating recidivism

risk, followed by the same question after specifying that the feature made the estimation more accurate, and finally after specifying that the feature made Black people more likely to be assessed as high risk. Their results show that, for all features, feature-disparity fairness is notably lower and feature-accuracy fairness is slightly higher than feature-apriori fairness.[8] Then, Grgić-Hlača *et al.* [40] use their feature-specific definitions to compare overall fairness levels of the $2^9 = 512$ different possible classifiers using subsets of the nine COMPAS features. They find that, as process fairness increases in classifiers that don't use the features labelled as unfair, accuracy levels decrease, but levels of outcome fairness (defined using Zafar *et al.* [88]'s disparate mistreatment) can remain high.

Grgić-Hlača *et al.* [40]'s work provides an interesting starting point for descriptively rooted definitions, but is reliant on an unwarranted assumption: that surveying the public will result in a consensus of what features are fair or unfair to use. In fact in their own later work, GH18 show that very low levels of consensus are reached for many of the features specific to the COMPAS dataset. In an attempt to come closer to some ground truth relating to fair feature selection, Van Berkel *et al.* [85] explore the levels of consensus that can be reached when crowdworkers are allowed to discuss their opinions together. Survey participants were assigned to groups of three, and then asked to vote whether or not specific features should be used by ADMs both before and after being allowed to discuss between group members. Analyses of the discussions show that participants considered each other's opinions, sometimes changed their minds, and overall made informed choices with the information they were given. Interestingly, they also find that groups that were more demographically diverse were more likely to come to an agreement in line with the majority. Van Berkel *et al.* [85] conclude that their method of crowdsourcing fairness perceptions is scalable and practical, but their results are not perfect and still leave multiple 'borderline' features that don't reach any consensus of fairness.

[8] Our results concerning the effects of accuracy on perceived fairness (Section 4.2) are in line with Grgić-Hlača *et al.* [41]'s feature-accuracy fairness results, but differ slightly. Their survey asks about the fairness of features *given* the fact that they increase accuracy, while our survey asks about a feature's perceived fairness and effects on accuracy *separately*.

2.2.3 Understanding Differences in Fairness Perceptions

Since attempts to discover an agreed upon truth about what it means for ADMs to be fair have been promising but mostly unsuccessful, an alternate approach is to instead attempt to understand how exactly it is that people perceive and think about fairness. The goal of understanding human judgement is by no means a novel one, and plenty of work has gone into breaking down the reasoning behind human judgement in other fields. For example, Graham *et al.* [37]'s *moral foundations theory* postulates that human morality is governed by a set of natural foundations including care, authority, and fairness. Konow [57] also proposes a similar idea with his *integrated justice theory*, which explains moral judgements using four different elements of justice.

Taking a page out of these earlier books, GH18 propose that people's moral reasonings about whether or not a particular feature is fair to be used by ADMs is governed by properties about the feature. Working again with features from COMPAS in the context of recidivism prediction, they propose that eight latent feature properties determine fairness judgements: *relevance, causes outcome, reliability, causes disparity in outcomes, caused by sensitive group membership, causes vicious cycle, volitionality,* and *privacy.* GH18 do not claim that their list of properties is exhaustive or complete, but they do claim to show that the list is both necessary and sufficient to describe their participants' reasonings. To demonstrate that these properties determine fairness judgements, AMT workers were asked to rate how fair they thought each feature is to use in the context of recidivism prediction, and also asked to rate how strongly they felt that each feature holds each property. Then, GH18 show that, despite little consensus being reached as to which features are fair or unfair to use, a participant's property assignments of each feature can be used to predict[9] their fairness judgement of that feature with $> 87\%$ accuracy. This remarkably high accuracy gives cause for optimism: that perhaps subjective disagreements about fairness judgements are actually caused by disagreements about the (arguably more objective) properties held by each feature, and that

[9]Using a logistic regression model with l2 regularization, randomly split into five 50%/50% train/test folds.

there is still a shared moral reasoning connecting the property assignments and the fairness judgements. GH18 conclude by remarking that if future work could determine objective assignments of which properties are held by each feature, then it could still be possible for consensus to be reached about the fairness of using individual features.

Another possible cause for disagreements in fairness judgements could be actual differences of opinion between people of different backgrounds. Pierson [70] demonstrates this by showing that women are significantly less likely to favour the inclusion of gender as a feature in a course recommendation ADM (despite increasing overall accuracy) if it resulted in fewer science courses being recommended to them. Grgić-Hlača *et al.* [39] similarly show that age, gender, and political leaning are significantly correlated with fairness judgements. Conversely, Wang *et al.* [86] find no significant effects of gender, education level, age, or race on perceived fairness, and Araujo *et al.* [3] find that age and gender have significant effects on perceptions of usefulness, but not perceptions of fairness. Considering these conflicting results, future work is required to continue to explore the effects of demographic differences on perceptions of fairness, especially pertaining to differences across racial groups, which are particularly understudied.

Finally, multiple studies have found that perceptions of fairness in algorithmic decision-making are dictated (and can be altered) by how well people understand the algorithmic processes that they are considering. For example, Pierson [70] shows that undergraduate students become more favourable to the use of ADMs in recidivism prediction after participating in an hour-long lecture discussing issues in algorithmic fairness. Explainability and transparency in algorithmic decision-making is often required by law [29, 36], but the type or degree of explanation provided to users can vary, and furthermore have varying effects on perceptions of fairness. Binns *et al.* [11] (and Dodge *et al.* [26]) explore these effects by comparing four styles of explanation that could be provided to the subject of a decision: *input influence*, presented as a list of features with their positive or negative quantitative effects on the outcome; *sensitivity*, presented as a list of features with the amount of change in value

for each feature that would have reversed the outcome; *case-based*, presented as a single case from the training data that is most similar to the current case and outcome; and *demographic*, presented as aggregate statistics on the outcomes of people in similar demographic groups. Both studies find significant differences in perceived fairness between explanation styles, specifically that the use of case-based explanations lowers perceptions of fairness. Additionally, qualitative responses highlight an interesting reason for preference of some explanations over others that Binns *et al.* [11] refer to as *actionability*: whether or not the explanation provides actionable recourse that a person could use to change their outcome in the future. For example, some participants perceived sensitivity style explanations as more fair when they listed features that could potentially be changed, like by increasing your income, as opposed to other participants that perceived demographic style explanations as unfair when they listed aggregate statistics of features that cannot be changed, like race. Ustun *et al.* [83] further advocate for the concept of actionability and provide a mathematical framework to ensure that ADMs are able to provide actionable recourse to the subjects of their decisions.

Understanding how perceptions of fairness are formed is a difficult task, as evidenced by the differing conclusions presented above. Unfortunately, the task is made even more difficult with the ever growing number of domains in which ADMs are used in real life. Even if we could determine an agreed upon definition of fairness in the context of recidivism prediction, that definition may need to change when applied to other contexts like loan or medical decisions. Worse, Araujo *et al.* [3] show that even within a single domain, perceptions of fairness can differ with varying levels of impact that a decision has on a person. There are likely multitudes of other unknown factors specific to certain people or contexts that also affect fairness perceptions (for example, Araujo *et al.* [3] also show that a person's level of concern about online privacy can affect how they perceive fairness). More research is required to carefully understand how fairness is differently perceived across different real-life decision-making contexts, and this thesis hopes to contribute to the continuing of that understanding.

2.2.4 The Effects of Fairness Perceptions

Partly prompted by Angwin *et al.* [2]'s controversial article, there has been a surge in work on fairness and discrimination in machine learning in recent years. However this surge may be limited for now to the academic world. In order to push for positive change in the actual decisions that affect real people every day, it is necessary to convince the businesses that use ADMs that they should care about their user's perceptions of fairness. Woodruff *et al.* [87] cite this goal as motivation for their work where they interviewed members of potentially marginalized communities (based on race and class) in their workshop-style study. They find that, although concepts of algorthmic fairness and discrimination were largely unfamiliar to their subjects, once explained, they were met with concern and overall negative feelings. Workshop participants had differing perceptions of who or what is at fault for existing unfairness, varying from a lack of diversity in programmers to news and media outlets, but, no matter where they thought unfairness came from, they still placed accountability on companies to provide solutions. Furthermore, Brown *et al.* [12] find that the use of ADMs amplifies existing distrust in child welfare systems. Marcinkowski *et al.* [64] also highlight a similar point in the context of university admissions by showing that perceived procedural and distributive fairness each have an effect on a student's likelihood of protesting or exiting a university, as well as the university's organizational reputation. Combined, these studies show that perceptions of fairness that users of an ADM have can affect their feelings about the companies or stakeholders behind the program.

2.2.5 Algorithmic versus Human Decision Makers

We have seen that perceptions of fairness can have an effect on levels of trust in a business, but another important related question is how much trust people have in ADMs themselves, especially when compared to levels of trust in equivalent human decision makers (HDMs). Multiple researchers have compared perceptions of fairness, trust, and usefulness between ADMs and HDMs, using both between- and within-subject study designs, and finding mixed re-

sults. Kennedy *et al.* [52] performed an experiment where participants were first prompted to make a prediction, and then offered the chance to change their prediction after receiving 'advice' in the form of a different prediction made by either a computer or a person. They find that participants are much more trusting of advice that came from a computer prediction, despite being informed that the computer predictor does not perform better than untrained humans. In a similar experiment, Logg *et al.* [62] show that advice labelled as from an algorithmic source is trusted more even when identical in value and accuracy to the advice (given to other participants) labelled as from a human source. Using a within-subject design, Marcinkowski *et al.* [64] also highlight a preference for ADMs by showing that their participants considered computer-made decisions to be more fair and more unbiased than equivalent human-made decisions (in the context of university admissions). Presenting results that are somewhat more intermediate in their between-subject experiment, Araujo *et al.* [3] find that perceptions of fairness are not significantly different between ADMs and HDMs, but perceptions of usefulness are significantly higher for ADMs than for HDMs in the domain of medical decisions (but not in the domains of justice or media decisions). Lee [59] performed an interesting experiment by dividing prediction tasks into the categories of 'mechanical tasks' (like assigning and scheduling work tasks for employees) or 'human tasks' (like hiring and evaluating employees). She finds levels of perceived fairness and trust to be equal across ADMs and HDMs for the mechanical tasks, but lower in ADMs than HDMs for the human tasks. Finally, Dietvorst *et al.* [25] demonstrate that people will exhibit *algorithm aversion* after seeing a computer model perform. In their between-subject experiment, participants were assigned to conditions where they either got the chance to watch a model make predictions (and receive feedback) ahead of time or did not, and then had to choose between using their own prediction or the model's predictions in a final incentivized prediction task. Participants who did not see the model perform were more likely to choose the computer-made predictions over their own, while participants who did see the model perform were more likely to choose their own predictions, despite the fact that the model

outperformed the human-made predictions most of the time. Dietvorst *et al.* [25] theorize that this result is due to the fact that people have a lower tolerance for errors made by computers than they do for worse errors made by humans. Combined, the results of the experiments described above are mixed, but lead us to two conclusions. First, perceptions of trust and fairness between ADMs and HDMs are context-dependent. Second, people may have a tendency to initially overestimate (and place high levels of trust in) the accuracy of computer-made decisions, but that trust could be diminished after seeing the computer predictors perform.

Chapter 3

Survey Design

We surveyed a total of 2157 AMT workers from November 2019 to June 2020. Participants were paid $5.00 USD (or $2.50 for the two pilot surveys). AMT has been found to produce high quality and honest results especially when paying higher amounts [50, 80]. Our respondents were 59% male, 40% female, 66% below 40 years of age and 34% over. 77% had completed some post-secondary education and 22% had not. 68% identified as Caucasian, 20% as Asian, 8% as Black, 6% as Hispanic, and 1% as Aboriginal. For a more detailed analysis of our demographic breakdown, see Chapter 5. For the initial surveys, we required that respondents be "master workers," a qualification given by AMT to signify that a worker has previously submitted a high amount of approved survey responses, but for the later surveys we switched to requiring an over 95% HIT (human intelligence task) approval rate with at least 1000 HITs completed to increase survey efficiency. Loepp and Kelly [61] find that non-master AMT workers with an approval rate over 90% and at least 100 HITs produced similar results to master workers. Additionally, we compare a small batch of workers with 1000 HITs and over 95% approval rates to master workers and find no significant differences between the groups' answers (using a Mann-Whitney U [MWU] nonparametric test with a Bonferroni corrected p-value threshold of 0.05).

Our initial survey asked 485 AMT master workers to rate the fairness of eight to ten features used in one of six decision-making domains on a seven-point Likert scale, as well as to rate how strongly eight properties that may

contribute to fairness reasoning correspond to each feature (also on a seven-point Likert scale). This structure is a replication of GH18's survey with the same goal of understanding moral reasonings used when making fairness decisions, but extended to our five new domains in addition to the bail domain considered by GH18. The order of the features, as well as which of the fairness questions or the property questions were asked first, was randomized. At the end of the survey, participants were asked to fill out their demographic information including age, gender, level of education, income, and ethnicity. See the supplementary materials for the exact survey wordings.

3.1 Domains and Features

Survey respondents were randomly assigned one of the six domains described below, which include GH18's original bail domain and five new domains. The new domains were chosen to cover a wide array of life-altering decisions that may be aided or made entirely by machine learning processes. Some of the domains are inspired by real machine learning programs and others are inspired by decisions that could benefit from the use of machine learning techniques. In all surveys, participants were given a description of the decision being made in their domain, and then asked to judge the fairness of each feature for that domain on a seven-point Likert scale from "Very Unfair" to "Very Fair".

- *Bail*: COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) is a machine learning program used in numerous American states to predict the chance that a defendant will reoffend if granted bail [24]. Information is obtained about the defendant from a questionnaire that they fill out. We consider the same ten features as GH18 which are used by COMPAS to output a risk score. They are: *current charges, criminal history, substance abuse history, stability of employment and living situation, personality, criminal attitudes, safety of their neighbourhood, criminal history of their friends and family, quality of their social life*, and *level of education*. For each feature, respondents were asked the question "How fair is it to determine if a person can be

released on bail using information about the/their [feature]?"

- *Child Protective Services (CPS)*: The county of Allegheny in Pennsylvania, U.S.A. has adopted a machine learning program called AFST (the Allegheny Family Screening Tool) to help CPS staff when deciding whether or not to "screen in" (that is, start an investigation based upon) tips that they receive [16, 21, 84]. We consider ten features that are used by this screening tool. These features are more vague than the ones that we consider in other domains because the specific features of the program are not publicly available; but it is specified that the program does not use race as a predictor. The features are: *demographics of the child victim (excluding race), CPS history of the child victim, juvenile justice history of the child victim, public welfare history of the child victim, demographics of the parents or other involved adults (excluding race), public welfare history of the parents or other involved adults, demographics of the alleged perpetrators (excluding race), public welfare history of the alleged perpetrators, CPS history of all individuals named in the referral,* and *behavioural health history of all individuals named in the referral.* For each feature, respondents were asked the question "How fair is it to determine whether or not a tip should be screened in to CPS using information about the [feature]?"

- *Hospital Resources*: PARR (Patients at Risk of Readmission) is a machine learning program that predicts the risk that a hospital patient will be readmitted within a certain number of days or months [5, 10]. The program is useful because it can help doctors decide how to best allocate the time and resources that they have available. We consider eight features that are used by this program: *age, gender, race, place of residence, hospital where they were treated, current hospital admission, number of emergency hospital discharges,* and *history of major health conditions.* For each feature, respondents were asked the question "How fair is it to allocate additional doctor care to prevent readmission using information about a patient's [feature]?"

- *Insurance Rates*: Insurance companies use machine learning programs to decide on the rates that they will charge a person to be insured. Companies often offer online services to get an instant quote by filling in preliminary information. We consider eight features that are referenced in online quote systems:[1] *age, gender, marital and family status, employment status, credit history, level of education, place of residence*, and *history of major health conditions*. For each feature, respondents were asked the question "How fair is it to determine an applicant's insurance rates using information about their [feature]?"

- *Loans*: When deciding whether or not to grant a loan to an applicant, banks use machine learning programs to predict the chance of loan default [23]. We consider ten features that may be used by banks to make loan decisions: *loan amount, income, age, gender, marital and family status, number of dependents, level of education, employment status, credit history*, and *owned property value*. For each feature, respondents were asked the question "How fair is it to determine an applicant's eligibility for a loan using information about the/their [feature]?"

- *Unemployment Aid*: In 2014, the Polish government introduced a machine learning program to classify citizens seeking unemployment aid into one of three categories that dictate the unemployment programs for which they are eligible [68]. We consider ten features that are obtained by this program either directly or through the unemployed person's responses to a questionnaire. They are: *age, gender, level of education, work history over the last five years, professional skills, degree of disability, time spent unemployed, place of residence, reason for wanting a job (other than income),* and *initiative*. For each feature, respondents were asked the question "How fair is it to determine a person's eligibility for unemployment aid using their [feature]?"

[1] The authors requested online instant quotes from a number of insurance firms in order to gather this list of features.

3.2 Properties

GH18 propose eight properties of features that contribute to a person's reasoning about fairness, which are listed below. In our initial survey, participants were given statements that assign each property to each feature and asked to rate how strongly they agreed using a seven-point Likert scale from "Strongly Disagree" to "Strongly Agree".

- *Relevance*: A feature is relevant if the information that it refers to is relevant to the decision being made. For each feature, respondents were given the statement "Information about the/their [feature] is relevant to [domain] decisions."

- *Causes Outcome*: A feature causes the outcome if the information that it refers to directly or indirectly causes the predicted outcome. For each feature, respondents were given the statement "The/Their [feature] can cause them to [domain outcome]" (for example, can cause them to recidivate).

- *Reliability*: A feature is reliable if the information that it refers to can be obtained reliably. For each feature, respondents were given the statement "Information about the/their [feature] can be assessed reliably."

- *Causes Disparity in Outcomes*: A feature causes disparity in outcomes if using the information that it refers to to make the decision may result in adverse consequences for people from certain protected groups. For each feature, respondents were given the statement "Making this decision using information about the/their [feature] can have negative effects on certain groups of people that are protected by law (e.g., based on race, gender, age, religion, national origin, disability status)."

- *Caused by Sensitive Group Membership*: A feature is caused by sensitive group membership if the information that it refers to is directly or indirectly caused by the person's belonging to a protected group. For each feature, respondents were given the statement "The/Their [feature]

can be caused by their belonging to a group protected by law (e.g., race, gender, age, religion, national origin, disability status)."

- *Causes Vicious Cycle*: A feature causes a vicious cycle if using the information that it refers to to make the decision may cause a cycle where the person is trapped in a pattern of risky behaviour or poor circumstances, disadvantaging them for future decisions. For each feature, respondents were given the statement "Making this decision using information about the/their [feature] can cause a vicious cycle."

- *Volitionality*: A feature is volitional if the information that it refers to is caused by choices made by the person involved. For each feature, respondents were given the statement "A person can change the/their [feature] by making a choice or decision."

- *Privacy*: A feature is private if accessing the information that it refers to requires a breach of privacy to the person involved. For each feature, respondents were given the statement "Information about the/their [feature] is private."

3.3 Pilot Survey: Framing Effects

To account for the difference that some domains and features were inspired by real machine learning programs (like PARR for the hospital domain) and others were postulated to be computer-aided (like the insurance domain), we completed a pilot study to test for framing effects introduced by the explicit mention of "machine learning". We asked 50 AMT master workers to rate the fairness of 16 features from four domains[2] where the description of the decision being made either explicitly mentioned "machine learning computer programs" or presented as a human-made decision and had no mention of computers or machine learning. We find that participant answers from the

[2]Respondents to this survey answered all questions instead of being assigned a specific domain, so only four domains were included to keep the survey brief. The domains were chosen to cover diverse scenarios and were: bail, hospital resources, loans, and unemployment aid.

two separate groups were not significantly different ($p > 0.05$) according to a Bonferroni-corrected MWU test. We conclude that no framing effects are introduced by the explicit mention of machine learning in decision-making.

3.4 Reproduction of GH18

Within the bail domain, our work is a reproduction of that done by GH18 and our results are consistent with theirs. We find similar levels of (lack of) consensus for both fairness judgements of the features and assignment of properties to features, as described in detail in Section 4.1.1. Using the property assignments, we are able to predict fairness judgements about the features with an accuracy of 82.6% in the bail domain (Section 4.1.2). This is lower than GH18's reported accuracy of 90.5% (using AMT data and excluding neutral responses), but still high.[3] The exact wordings of the Likert scale questions, described in Sections 3.1 and 3.2, are all inspired by GH18's original wording in the bail domain.

3.4.1 Pilot Survey: Necessity and Sufficiency

We also replicated GH18's necessity and sufficiency study (their pilot survey 1). GH18 claims that their eight proposed properties are both necessary to and sufficient for moral reasoning about fairness judgements of the features, and we replicate their method and results in the bail domain as well as in our five new domains. 425 AMT master workers were randomly assigned a domain and asked to rate the fairness of each feature for their domain on a Likert scale. However after each rating, instead of answering a Likert question for every property, respondents were asked to explain their fairness rating by selecting any number of the eight properties[4] and/or by filling in a text response. We

[3] In an effort to explain this difference in accuracy levels, we ran our analysis on GH18's AMT data, but achieved a slightly higher accuracy of 90.9% indicating that the issue was not with our analysis. It is likely that our data was noisier, which is supported by higher variance in accuracy over the 1000 train/test splits when using our data rather than using GH18's. An anonymous reviewer suggested that this could possibly be due to differences in the AMT worker populations between GH18's study in 2018 and our study in 2020.

[4] Properties were listed as explanations in sentence form, for example: 'The [feature] can/cannot be assessed *reliably*' (can or cannot used depending on whether the participant

find, similarly to GH18's bail domain, that within each domain every property is selected as an explanation by at least 30% of respondents and fewer than 9% of respondents used the fill-in the blank option (the majority of which were reiterations of the listed properties). Thus according to GH18's analysis method, the eight properties are both necessary, as a substantial fraction of respondents appealed to them in their explanation, and sufficient, as the fill in the blank option was left unused by over 90% of respondents, to *explaining* fairness judgements. Although, as we will see later (Section 4.1.3), this is not equivalent to being necessary for *predicting* fairness judgements.

3.5 Additional Surveys

Based on results from our initial survey (Section 4.1.3), we decided to complete two additional surveys with slight changes. These surveys were completed by AMT workers with at least 1000 HITs completed and an over 95% approval rate. First, 585 AMT workers completed a new version of the study where the property *relevance* was replaced by a new property we called *increases accuracy*:

- *Increases Accuracy*: A feature increases the accuracy of the predicted outcome if using the information that it refers to to make the decision increases the proportion of cases in which the correct decision was made. For each feature, respondents were given the statement "Using information about the/their [feature] would increase the accuracy of [domain] decisions."

Second, 592 AMT workers completed a version of the study where we removed *relevance* altogether from the list of properties. Aside from the property *relevance* being replaced or removed, all other parts of the surveys are identical to the previously described survey.

chose a 'fair' or 'unfair' option in the previous question). See the supplementary materials for more details.

Chapter 4

Analyses and Results

When we began this work, our initial goal was to test whether GH18's finding that their eight properties act as moral determinants for fairness judgements extends to other decision-making contexts, but also to test if moral reasonings are shared between contexts, and further understand which properties best determine fairness judgements. Section 4.1 outlines how we find low levels of consensus in all domains, but are able to predict fairness judgements well using the property assignments in every domain and across domains.

Surprisingly, we find that a participant's assignment of the property *relevance* to a feature is able to predict fairness judgements well entirely by itself, better than any combination of the other features (in most domains).[1] This result prompted us to complete two additional surveys with the purpose of better understanding how participants had interpreted the property *relevance*. Section 4.2 outlines our findings that the property *relevance* behaves quantitatively similarly to the property *increases accuracy*, which we conclude to be the most important property used in moral reasoning about fairness judgements (by our participants).

4.1 Generalization Across Domains

In our initial analysis, we set out to answer three questions. First, is the lack of consensus found by GH18 repeated in all six domains? Following GH18,

[1] This result is true in all domains when using all of our data, but is not true for the hospital and insurance domains after removing Caucasian and male respondents, respectively. See Section 5.2 for more details.

we use Shannon entropy to calculate the degree of consensus in fairness judgements of the features and assignment of properties to features. Second, can fairness judgements be accurately predicted using property assignments within and across the six domains? We train logistic regression fairness classifiers to show the predictive power of the properties.[2] Third, we ask which of the properties are the most predictive of fairness judgements, which we answer in two ways: by comparing the weights assigned by each property in the models, and by comparing the accuracies between models using subsets of the properties (especially using only a single property). Whenever comparing two groups, we use the Mann-Whitney U (MWU) test and apply the Bonferroni correction before determining significance, as our data is not necessarily normally distributed.

4.1.1 Consensus

We find high levels of disagreement across both fairness judgements of features, and assignment of properties to features. However, we also find a strong positive correlation between the consensus in which properties apply to a feature, and the consensus of how fair it is to use a feature. Following GH18, we determine levels of consensus using Shannon entropy [75]. Shannon entropy measures the uncertainty or surprise in a random variable's possible outcomes using the average number of bits required to describe the possible outcomes. Higher entropy corresponds to lack of certainty in outcomes (for example when flipping a fair coin) and lower entropy corresponds to more certainty (for example when flipping a biased coin). This value can also be interpreted as a measurement of consensus in survey responses instead of certainty over random variable outcomes; in fact, Shannon entropy has been commonly adapted to quantify consensus in Likert scale questions [81]. Specifically, we calculate the Shannon entropy over the proportions of responses bucketed into three

[2]We use logistic regression specifically as a replication of GH18's work. However, we also tested five other classifiers (support vector classification [SVC], linear SVC, k-nearest neighbors, decision trees, and random forests) and found that only linear SVC obtained similar accuracies to logistic regression, while the other classifiers under-perform logistic regression. Our qualitative conclusions are identical using linear SVC or linear regression.

categories: "Unfair", "Neutral", and "Fair", and then normalize this value to be between 0 and 1 by dividing by $\log_2 3$ (since that is the maximum possible entropy for a distribution over 3 values). We report 1 minus this normalized Shannon entropy (so that 1 corresponds to complete consensus and 0 to complete disagreement) for each of the feature's one fairness judgement question and eight property assignment questions (all values are presented in Table A.1 in the supplementary materials).

Observing only the fairness judgement questions, we find that, with the exception of the loan domain, little consensus in fairness judgements of the features is reached for most features of all domains. The higher levels of consensus in the loan domain are possibly due to our inclusion of multiple features that relate directly to finance, including income and employment status, especially since other features like age still achieved lower consensus. Similarly to GH18, we find that only current charges and criminal history achieve high consensus in the bail domain. In fact out of the 56 features across all domains, only 10 features achieve levels of consensus above 0.5 (five of those are in the loan domain) and 30 features reach levels of consensus lower than 0.3.

The assignments of properties to features result in even higher levels of disagreement. Out of the 448 combinations of properties and features about which participants were asked (8×56), only 44 achieve a consensus level greater than 0.5. The highest consensus is achieved when rating how relevant or reliable a feature is (which echoes findings by GH18), with some higher levels for volitionality and causes outcome as well. After averaging across the eight properties to obtain a single measure of consensus per feature, the features with the highest average agreement in property assignments are again current charges and criminal history in the bail domain, as well as age in the insurance domain and a few features in the loan and hospital domains. Comparing the levels of consensus in fairness judgements with the levels of consensus in average property assignments for each feature reveals a very strong positive Pearson correlation coefficient of 0.72, supporting the idea that the assignments of properties to features act as moral determinants to fairness judgements of the features.

Domain	Within	Cross Trained	Cross Tested
Bail	$82.6 \pm 0.00047\%$	$80.8 \pm 0.015\%$	$81.4 \pm 0.006\%$
CPS	$82.7 \pm 0.00047\%$	$78.7 \pm 0.018\%$	$80.5 \pm 0.011\%$
Hospital	$78.8 \pm 0.00057\%$	$81.7 \pm 0.015\%$	$78.4 \pm 0.005\%$
Insurance	$80.5 \pm 0.00054\%$	$80.0 \pm 0.018\%$	$79.6 \pm 0.012\%$
Loan	$87.6 \pm 0.00042\%$	$80.2 \pm 0.008\%$	$86.4 \pm 0.004\%$
Unemployment	$77.7 \pm 0.00050\%$	$82.4 \pm 0.014\%$	$77.4 \pm 0.005\%$
All (Pooled)	$82.0 \pm 0.00019\%$	N/A	N/A

Table 4.1: The average accuracies and standard errors of the mean for the within-domain, all-domain, and cross-domain predictors (predicting fairness judgements using property assignments). *Within:* Accuracies obtained by training and testing the model within the listed domain (or over all pooled domains), each average is over 1000 50%/50% train/test splits. *Cross Trained:* Accuracies obtained by training in the listed domain and testing in each of the other domains individually, average is over the five other domains used to test. *Cross Tested:* Accuracies obtained by training in each of the other domains individually and testing in the listed domain, average is over the five other domains used to train.

4.1.2 Predicting Fairness Judgements

If the properties act as moral determinants for fairness judgements, then we should be able to use machine learning algorithms to predict respondents' fairness judgements for a given feature using their assignment of properties to that feature. To test this, we train multiple logistic regression classifiers using l2 regularization to predict whether a subject rated a feature as unfair (answered "Very Unfair", "Unfair", or "Somewhat Unfair") or fair (answered "Very Fair", "Fair", or "Somewhat Fair"). GH18 finds higher accuracy when excluding "Neutral" responses from their data, possibly because respondents who selected "Neutral" did not have clear moral reasonings for their selection, so we choose to exclude those points as well. We perform 1000 50%/50% train/test splits within each domain (individually), as well as over all domains (pooled). Table 4.1 (first column) reports the average accuracy for each predictor. The individual domain accuracies range from 77.7% in the unemployment domain to 87.6% in the loan domain. In the bail domain, we find an accuracy of 82.6%. When pooling all domains together, we achieve a remarkably high accuracy of 82.0%. The fact that our pooled domain accuracy is compara-

ble to our individual domain accuracies (and even higher than the individual domain average of 81.6%) suggests that the properties hold similar relative importances across domains.

In order to further investigate how the properties transfer between domains, we also perform cross-domain tests where a classifier is trained in one domain and tested in another. For all six domains, we train a model using that domain's data and test it on each of the five other domains' data. Table 4.1 (second column) presents each domain's accuracy when used for training and averaged over the other five testing domains. Additionally, Table 4.1 (third column) presents the same results averaging instead over the five other domains used for training while testing in the listed domain. In all, respondents' property assignments are able to well predict their fairness judgements regardless of the decision-making domain that was used for training the model. This leads us to conclude that, despite lack of consensus in fairness judgements of the features, our respondents used similar moral reasonings when making fairness decisions across the different domains.

4.1.3 Predictive Power of Relevance

We next investigate the relative predictive power of each property. First, we observe the weightings associated with the properties by domain (Figure 4.1 top), which reveals that only *relevance* has a consistently significant effect on fairness predictions.[3,4] *Causes outcome* has a significant effect in the bail and hospital domains, but no other properties have significant weights in any of the domains. Second, we train new predictors using subsets of the eight properties, including ones using only a single property. Remarkably, we find that the models using *relevance* as their sole predictive feature frequently perform as well as (or better than) their complete eight-property equivalents, while many of the other seven single-property models barely outperform a baseline, if at all (Figure 4.2). In fact, with some exceptions (see Section 5.2), no combina-

[3]Significance tests performed using a linear regression t-test with the null hypothesis that the coefficient values are equal to zero.

[4]With the exception of the insurance domain after removing male participants from our data, see Section 5.2 and the supplementary materials for more details.

tion of properties excluding relevance within any of the domains or the pooled domain models is able to predict as well as *relevance* alone in the equivalent model.[5] These results also hold in GH18's original data. Why would *relevance* be able to predict so well on its own? After review of our survey wording, we hypothesized that the word "relevant" may have been interpreted in one of two ways. First, it is possible that respondents interpreted "relevant" as a synonym for "fair" and so agreed that any feature which they thought was fair was also relevant. Second, respondents may have interpreted the word "relevant" to mean that a feature would increase the accuracy of the decision being made. The next section outlines the additional surveys that we completed to investigate our respondents' interpretation of relevance.

4.2 Beyond Relevance

In the initial survey, participants were asked to rate how much they agreed that information about specific features was relevant to the decision being made. Their answer to that question alone often turned out to be highly predictive of how fair they thought it was to use said information in decision-making processes. In trying to understand why relevance as a property is able to predict fairness judgements of the features so well, we determine two possible ways that relevance could have been interpreted by the initial survey respondents:

- A feature was labelled as relevant to the decision if it would increase decision accuracy
- A feature was labelled as relevant to the decision if it was considered fair to use in decision-making

[5]These results might initially seem to contradict our pilot survey that finds all eight properties to be necessary. In fact we believe they are consistent. Our pilot survey determines that all properties were used by respondents to *explain* their own fairness judgements. These results indicate that only *relevance* is needed to *predict* fairness judgements of the features. The other properties may in fact have formed part of the respondents' reasoning. Respondents could select multiple properties as part of their explanations, so some properties' being selected only in tandem with *relevance* would have produced exactly this combination of necessity and predictiveness.

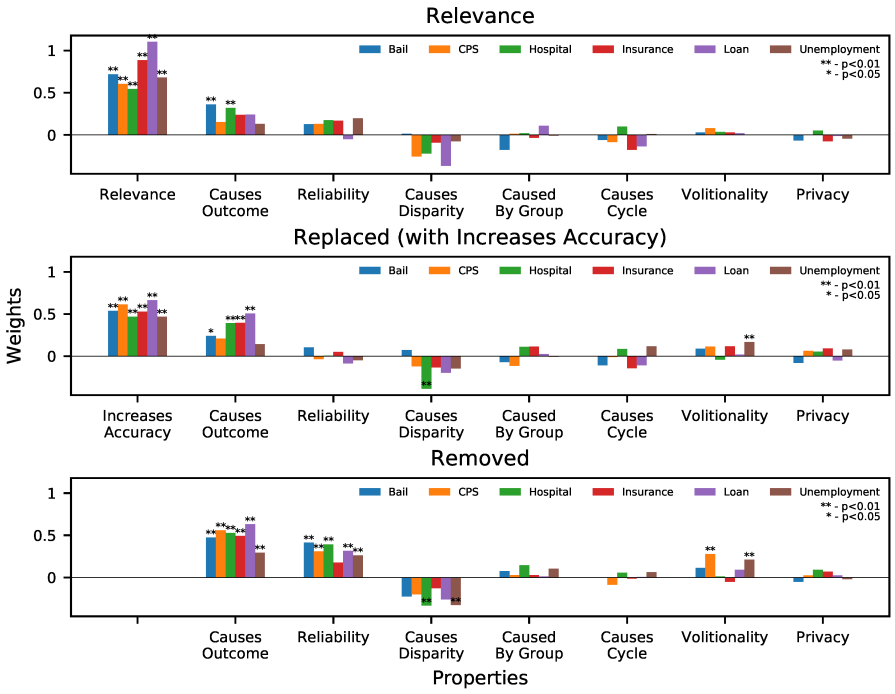

Figure 4.1: The weights associated with each property in every domain for the full property models of the initial survey (Relevance), the survey that replaced *relevance* with *increases accuracy* (Replaced), and the survey that removed *relevance* all together (Removed). Significant weights (using a linear regression t-test with the null hypothesis that the coefficient values are equal to zero) at the $p < 0.05$ (*) and $p < 0.01$ (**) levels are indicated.

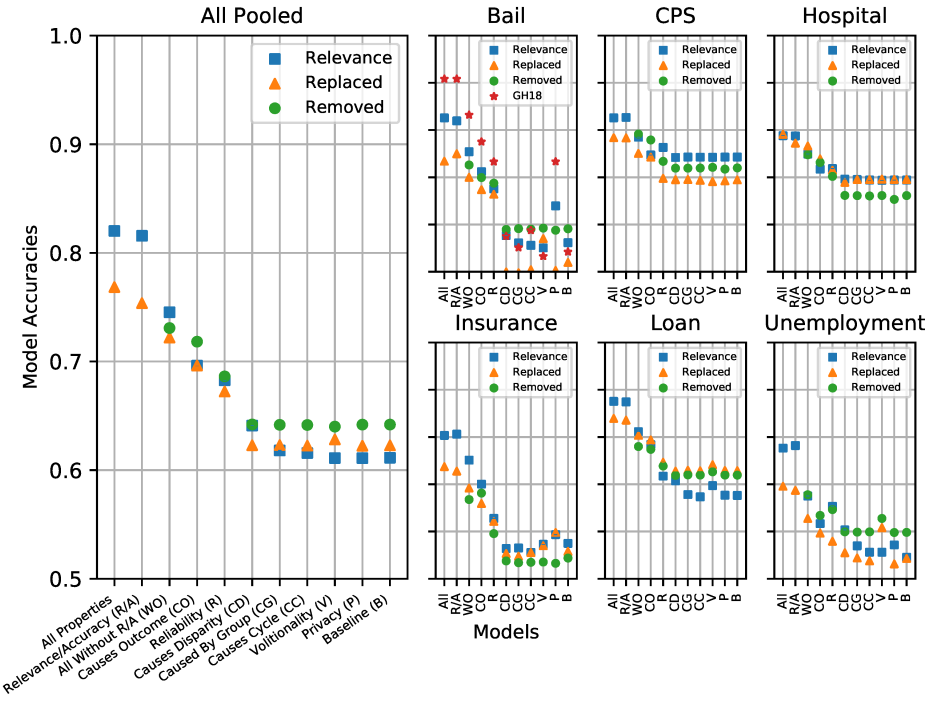

Figure 4.2: The accuracy levels achieved per model. For each domain (and all pooled), the accuracies achieved from the initial survey (Relevance) and additional surveys (Replaced and Removed) are shown for the following models where applicable: using all eight properties, using only *relevance/increases accuracy*, using all seven properties excluding *relevance/increases accuracy*, and using each of the other seven properties individually. Additionally for comparison, we show the baseline accuracies that are achieved by guessing only either "Fair" or "Unfair" (whichever does better). Error bars are present for each point but are too small to be visible.

To gain a better understanding of how participants interpreted the concept of relevance, we designed two new surveys with slight alterations to the original.

4.2.1 Replacing Relevance with Accuracy

For our first altered survey, we replaced *relevance* with a more specific property that is less open to differing interpretations:

- *Increases Accuracy*: A feature increases the accuracy of the predicted outcome if using the information that it refers to to make the decision increases the proportion of cases in which the correct decision was made. For each feature, respondents were given the statement "Using information about the/their [feature] would increase the accuracy of [domain] decisions."

If *relevance* had been interpreted as *increases accuracy*, then we would expect to see similar qualitative patterns in the new survey results as in the initial survey results. On the other hand, if *relevance* had been interpreted as a synonym for fairness, then we would expect prediction accuracy to be lowered by replacing *relevance* with *increases accuracy*. Results from the new models using this survey are shown in Figure 4.2 (as "Replaced"). With the exception of the hospital domain, all full property "Replaced" models perform worse than their initial survey ("Relevance") full property model equivalents, indicating that at least some original survey respondents may have interpreted *relevance* as a synonym for fairness. This conclusion is also supported by the slight increase in the total number of other properties with weights that significantly differ from zero, shown in Figure 4.1 (top to middle). However, for most domains, fairness judgements are still well predicted by the new *increases accuracy* property on its own.[6] Moreover, the replaced property is still better at predicting fairness than any combination of the other properties in most cases, and the *increases accuracy* weightings in the eight-property models are mostly still consistently higher than any other property. This

[6]This is true with the exception of the hospital and insurance domains after removing Caucasian and male participants respectively, see Section 5.2 for more details.

repeated pattern of results is supportive of our hypothesis that many initial survey participants labelled a feature as relevant if they thought its inclusion would increase decision accuracy. Although, we will see in Chapter 5 that these results differ for some domains after removing privileged groups from our data.

4.2.2 Removing Relevance

Compared to *relevance/increases accuracy*, none of the other properties are close to as highly predictive of fairness judgements. However, because our evidence is consistent with some participants having interpreted *relevance* as a synonym for fairness, it is possible that the existence of *relevance* as a listed property reduced the predictive power of the other properties. Survey respondents may have so strongly equated *relevance* to fairness that their other property assignments would be given less magnitude. To check for this, our second altered survey removed *relevance* as a property, without replacing it with a new property. If the inclusion of *relevance* had reduced responses to the other properties, then we would expect to find an increase in model accuracy when compared to the model using original survey results without *relevance*. Results are presented in Figure 4.2. The only domains that showed an increase in model accuracy were unemployment and CPS, but both increased by less than 0.01%. However, Figure 4.1 (bottom) shows that the removal of *relevance* did significantly increase the weightings for *reliability* (for all domains except insurance) and *causes outcome* (for all domains), as well as *volitionality* and *causes disparity in outcomes* for a few domains. This finding indicates that the inclusion of *relevance* as a property may have somewhat distracted from the other properties, but not enough to create a noteworthy difference in prediction accuracy. Overall, we conclude that *relevance* (or as most respondents seem to have interpreted it: *increases accuracy*) is the most important property used in moral reasoning about fairness judgements by our survey respondents.

Chapter 5

Demographic Analyses

A limitation to our study is the lack of diversity in our sample: our survey respondents were primarily younger (<40 years old), male, Caucasian, and a large majority had completed some form of post-secondary education. This is especially problematic in the context of life-affecting algorithmic decision-making because the people most likely to be adversely affected by ADMs are those from minority and often under-represented groups [2, 7]. However, due to our large total number of participants across all surveys, we still have enough members of most demographic groups to be able to perform demographic analyses. For example, Black participants make up only 8% of our total sample population, while making up 14% of the American population [82]. But that means that an American census-representative sample of 200 people would include about 28 Black participants, which is fewer than the 34 Black participants that we have solely in the unemployment domain. In other words, the total number of Black participants that we have just in the unemployment domain would be representative of the United States population in a sample of 243 people (Table 5.1) which is comparable to sample sizes used in other works [e.g., 26, 39, 40, 59, 74, 79, 85]. This argument can be repeated: for most demographic groups, we have more total respondents per domain than an American census-representative sample of 100 people would (see Table A.2 in the supplementary materials for all demographic groups).

Despite our large total numbers of participants, the problem remains that our results could be overrepresentative of the majority groups in our sample,

Ethnicity	Bail	Unem.	CPS	Hos.	Loan	Ins.
Black	214	243	136	143	157	186

Table 5.1: The hypothetical sample sizes for which our total number of Black respondents per domain would be representative of the United States population [82], rounded to the nearest whole number. See Table A.2 in the supplementary materials for all demographic groups.

but, *because* of our large total numbers of participants, we are able to perform demographic analyses to investigate if this is the case. We do so in two ways. First, we create new logistic regression models to determine if any of the demographic features have significant effects[1] on the fairness judgements or the property assignments. Second, we remove the privileged demographic groups that make up the majority of our data (Caucasians and males) and repeat our main analyses to search for differences in results. The results of these two demographic analyses are outlined in the next sections.

5.1 Significant Effects of Demographic Information

Inspired by Grgić-Hlača *et al.* [39], we check for significant effects of demographic information on the fairness judgements using a new logistic regression model identical to the one described in Section 4.1.2, but with the independent variables being demographic information[2] instead of property assignments. Additionally, we complete the same analysis with each of the property assignments as the dependent variables. The resulting weights that are significantly different from zero are shown in Table 5.2 and Table 5.3, presented as Bonferroni corrected 95% confidence intervals. Overall, we find very few significant demographic effects. Out of the 504 fairness judgement predictor weights (9 demographic features $\times$ 56 fairness judgements) and the 4032 property assignment predictor weights (9 $\times$ 56 $\times$ 8 properties), only 5 and 16 weights, respectively, were significantly different from zero. Interestingly, with the ex-

[1]Significant effects meaning that the coefficients are significantly different from zero.

[2]Age, level of education, and income are expressed using integers for each categorical value and gender and each separate ethnicity group are treated as binary "dummy variables".

Domain	Feature	Demo. Feat.	95% C.I.
Bail	Current Charges	Black	0.34 - 1.77
Hospital	History of Health	Black	0.22 - 1.69
Loan	Income	Black	0.20 - 1.73
Loan	Gender	Education	0.03 - 1.33
Loan	Employment	Black	0.11 - 1.54

Table 5.2: The weights of the demographic features that are significantly different from zero when used to predict fairness judgements, using logistic regression, presented as Bonferroni corrected 95% confidence intervals. The intervals are obtained by bootstrapping with replacement 1 million times. Only these five of the 56 total fairness judgements have a single demographic feature weight (out of nine weights each) that is significantly different from zero. Interestingly, all of the significant weights are positive, meaning that members of that demographic group were more likely to think that that feature is fair (rather than unfair). This could be due to lower total numbers of respondents that thought a feature was unfair in general, resulting in less total data points for significance testing with unfair features.

ception of level of education in the loan domain (which has a significant effect on the perceived fairness of using gender as a feature), all of the significant effects are from ethnicities. Additionally, out of the five features with fairness judgements that are significantly affected by demographic information, only two of them (income and employment status in the loan domain) have equivalent property assignments that are significantly affected by demographic information. These results together tell us that: not only do differences in fairness judgements across demographic groups exist (although sparsely), they are seemingly not introduced by differences in property assignments. Determining the root causes of these demographic differences of opinion, for both fairness judgements and property assignments, is an important direction for future work.

5.2 Removing Privileged Groups

Next, we repeat our main analyses twice: first using data without any Caucasians, then without any males. Full results from these analyses are presented in the supplementary materials. Initial results from both analyses introduce no changes to our early conclusions, since GH18's eight properties are still able

Domain	Feature	Property	Demo. Feat.	95% C.I.
Bail	Criminal History	Reliability	Black	0.29 - 1.62
CPS	Child CPS His.	Causes Outcome	Hispanic	0.14 - 1.51
CPS	Juvenile Justice	Causes Outcome	Black	0.34 - 1.93
CPS	All CPS His.	Causes Outcome	Hispanic	0.24 - 1.74
CPS	All Health His.	Causes Outcome	Hispanic	0.36 - 1.84
Hospital	Place of Residence	Privacy	Hispanic	0.08 - 1.53
Hospital	No. of Discharges	Reliability	Hispanic	0.04 - 1.39
Hospital	No. of Discharges	Relevance	Asian	0.05 - 1.62
Insurance	Employment	Volitionality	Hispanic	0.20 - 1.64
Insurance	History of Health	Privacy	Hispanic	0.17 - 1.64
Insurance	History of Health	Causes Cycle	Hispanic	0.35 - 2.13
Insurance	History of Health	Causes Disparity	Hispanic	0.37 - 2.17
Loan	Loan Amount	Reliability	Black	0.18 - 1.68
Loan	Income	Causes Outcome	Black	0.07 - 1.65
Loan	Employment	Volitionality	Black	0.25 - 1.64
Loan	Employment	Causes Outcome	Black	0.12 - 1.61

Table 5.3: The weights of the demographic features that are significantly different from zero when used to predict property assignments, again using logistic regression, and again presented as Bonferroni corrected 95% confidence intervals obtained by bootstrapping with replacement 1 million times. Only these 16 of the 448 total property assignments (56 features $\times$ 8 properties) have a single demographic feature weight (out of nine weights each) that is significantly different from zero. All of the significant weights are again positive, now meaning that members of that demographic group were more likely to think that a feature does have a property, which is again possibly due to lower total numbers of respondents in general who think a that a feature does not hold a property.

to well predict fairness judgements both within and across domains. Additionally, in *most* domains, both analyses introduce no changes to our conclusion that *increases accuracy* is the most important property for predicting fairness judgements. However, observing the predictive powers of the properties in the models without Caucasians and without males each reveal surprising differences in two of the six domains: hospital resources and insurance rates.

Using data without Caucasian participants, results in the hospital domain indicate that *relevance* and *increases accuracy* no longer have significant effects on fairness judgements. In fact, across all surveys excluding Caucasians, the only properties with significant effects on fairness judgements in the hospital domain are: *causes outcome* in the survey that included *increases accuracy*; and *causes outcome* and *reliability* in the survey that removed *relevance* entirely. However, despite no longer having a significant effect on fairness judgements in the full eight-property model, the single property model using only *relevance* still performs better than any combination of the other properties. Additionally, in the survey that included *increases accuracy*, the single property model using *causes outcome* also now performs better than all combinations of the other seven properties, instead of *increases accuracy* ("Replaced" in Figure 5.1, left). These results suggest that our non-white participants in the hospital domain may have interpreted *relevance* as more of a synonym for *causes outcomes* than *increases accuracy*, although overall these results are difficult to interpret within the frame of our other conclusions and warrant future research.

Using data without males, results in the insurance domain reveal that *increases accuracy* no longer follows the same qualitative pattern as *relevance*. The full eight-property predictive model using results from the survey with *increases accuracy* now performs significantly worse than that using results from the initial survey with *relevance* ("Replaced" in Figure 5.1, right). Furthermore, the only property that now has a significant effect on fairness judgements in any of the surveys (except for *relevance* in the initial survey) is *causes outcome* in the survey that removed *relevance* as a property. These results could be indicative that female participants in the insurance domain of our initial

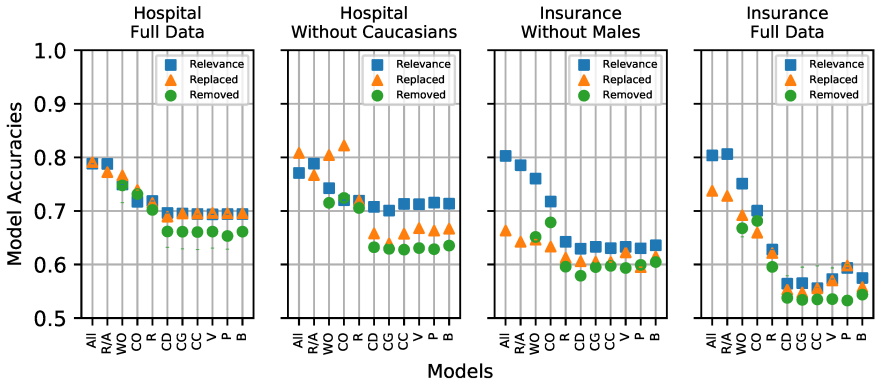

Figure 5.1: An identical subset to Figure 4.2, but using data without Caucasians in the hospital domain and without males in the insurance domain. Figure 4.2 caption: The accuracy levels achieved per model. For each domain (and all pooled), the accuracies achieved from the initial survey (Relevance) and additional surveys (Replaced and Removed) are shown for the following models where applicable: using all eight properties, using only *relevance/increases accuracy*, using all seven properties excluding *relevance/increases accuracy*, and using each of the other seven properties individually. Additionally for comparison, we show the baseline accuracies that are achieved by guessing only either "Fair" or "Unfair" (whichever does better). Error bars are present for each point but are too small to be visible.

survey interpreted *relevance* more as a synonym for fairness than a synonym for *increases accuracy*. Overall, no property other than *relevance* emerges as highly important to moral reasonings about fairness judgements by our female participants in the insurance domain.

Outside of the hospital and insurance domains, results from our analyses after removing Caucasian and male participants from our data do not introduce any changes to our conclusions. This is also true for the pooled-domain analyses, in which the single property model using *increases accuracy* still performs better than all combinations of the other seven properties.[3] However, the differences that we find in the hospital and insurance domains indicate that future work should focus specifically on these domains while oversampling minority groups.

Regardless of the specifics of which particular domain features exhibit significant differences in fairness judgements or property assignments, the very ex-

[3]This fact remains true for further broken down subsets of data including intersectional combinations of race and gender.

istence of these differences across demographic groups, especially race, speaks to the importance of considering minority group opinions. This importance is also highlighted by previous work that finds negative reactions to ADMs by minority and adversely-affected groups [87]. Within the descriptive literature on algorithmic fairness, it is common practice to use American census-representative (but majority Caucasian) population samples, sometimes without any demographic-specific analyses [e.g., 38, 44, 74, 79, 85]. Due to the combined facts that minority populations' algorithmic fairness judgements may significantly differ from the majority, and that minorities are more likely to be adversely affected by ADMs, this practice is potentially harmful. In this study, we chose to prioritize having a large sample population over enforcing diversity, but also necessarily complete demographic analyses to identify when majority opinions are overrepresented. However, future descriptive work in algorithmic fairness could alternatively oversample minority populations in order to further understand when and why fairness judgements differ across demographic groups.

Chapter 6

Conclusion

In this work, we consider how disagreements in fairness judgements about machine learning-based decision procedures can be explained by a single set of properties across different domains. We consider six domains: bail, child protective services, hospital resources, insurance rates, loans, and unemployment aid. We descriptively analyze answers to two kinds of survey questions. The first asks participants to report fairness judgements about features, using questions of the form "How fair is it to [make a decision about a person] using information about the/their [feature]?" The second asks participants to rate how strongly they agree or disagree that a feature holds a specific property. Using these property assignments, we are able to successfully well predict fairness judgements both within and across domains as well as over all data pooled together.

Our initial conclusion is that GH18's eight feature properties do indeed describe the moral determinants used for making fairness judgements in all six domains. Furthermore, these properties are predictive in similar ways in each domain.[1] This is evidenced by our high prediction accuracy from our model using all domains' pooled data (82.0%), which outperforms the average of the individual within-domain predictors (81.6%). This conclusion is also supported by our cross-domain predictors which perform well after training within one domain and testing in another. Much of the literature in fairness related machine learning research has focused only on a single domain,

[1] That is, properties that strongly predict that a feature will be judged fair in one domain make the same prediction in the other domains.

especially the bail domain [e.g., 9, 15, 17, 38, 40, 41, 46, 79]. Our results thus provide confidence that existing work, especially other process-based fairness work, could be applied to other domains and scenarios.

Our most surprising result is that the property which performs best at predicting the perceived fairness of a feature, in most cases, is whether or not a feature would increase decision accuracy. Each of the two near-synonym properties *relevance* and *increases accuracy* predict fairness judgements of the features better when used individually than *any other* combination of GH18's eight properties, in most domains (with the exceptions of the hospital and insurance domains after removing Caucasian and male respondents, respectively, from our data). Furthermore, removing *relevance/increases accuracy* as a possible answer for property assignment questions does not substantially increase the predictive power of any other properties. This result is especially important because fairness-related machine learning research often references an "unavoidable" accuracy-fairness trade-off [e.g., 8, 40, 41, 66, 90]. In contrast, our results point to an important correlation between accuracy and fairness (specifically perceived process fairness). Grgić-Hlača *et al.* [40] and Grgić-Hlača *et al.* [41] find that process- and outcome-based fairness can be achieved simultaneously but at the cost of lowering decision accuracy. Our work, on the other hand, suggests that the level of decision accuracy could itself be used as a process-based fairness metric; in fact, decision accuracy has been previously studied as an indirect fairness metric [40, 41]. Error rates, which are an equivalent opposite measure to decision accuracy, have frequently been proposed and used as measures for outcome-based fairness [9, 43, 88], but usually only by equalizing them across protected groups. Similar to our results, Harrison *et al.* [44] also find that accuracy levels affect fairness judgements by comparing levels of fairness to bias. They determine that models can be perceived as non-biased but still unfair if error rates are equalized across protected groups but considered too high.

In their introduction, GH18 motivate their goal of determining the perceived fairness of individual features by referencing their previous work [40, 41], which proposes a definition for process fairness that becomes more fair

with the removal of certain features from the predictive model. However, the removal of unfair features also necessarily *decreases* decision accuracy levels; our results suggest that this would lead to a corresponding *decrease* in perceived fairness. In fact, earlier in the same work, Grgić-Hlača *et al.* [41] show that features are perceived as *more* fair when they are framed to increase decision accuracy, presenting a result that is more in line with our findings. In a similar vein, Kleinberg and Mullainathan [55] show that any "simple" model (for example using fewer predictive features) can be strictly improved in both efficiency and equity by replacing it with a more complex model. Of course, some features may simultaneously increase decision accuracy and cause negative impacts to protected groups of people (for example, due to unequal true base rates of recidivism among Black and white defendants caused by historical discrimination), which is why it is important to consider both process-based and outcome-based fairness measures when designing ADMs.

Because of our choice to prioritize a large sample size rather than enforce diversity, there was a risk that our results would be overrepresentative of privileged majority populations. In fact, this turned out to be somewhat true in the hospital and insurance domains, revealed by our two repeat analyses after removing Caucasian and male respondents from our data, but only pertaining to our conclusions about the predictive power of specific properties. In the hospital domain, using data without Caucasians, *causes outcome* is the most predictive property instead of *increases accuracy* in the survey where *relevance* was replaced by *increases accuracy*, but no properties have significant effects on fairness judgements in the initial survey with *relevance*. In the insurance domain, using data without males, no single property emerges as the most predictive other than *relevance*. We also find that only a small number of fairness judgements or property assignments of features are significantly affected by participant demographic information. Of the few fairness judgements and property assignments that are significantly affected by participant demographics, the majority relate to participant ethnicity. Furthermore, significant demographic effects on fairness judgements are not necessarily paired with effects on property assignments of the same feature, indicating that demo-

graphic differences in fairness judgements are not fully explained by differences in property assignments. Future work is needed (particularly in the hospital and insurance domains) to understand when and why people of differing demographic backgrounds, especially pertaining to race, have differing perceptions of fairness.

Our results suggest that the lack of consensus in fairness judgements can be partially explained by a lack of consensus in property assignments, but the lack of consensus about the property assignments themselves is yet to be understood. Factors such as framing effects or foundational values may play a role in property assignments; teasing apart their effects is another exciting future direction. Finally, as mentioned above, increases to decision accuracy, while improving perceived process fairness, may also diminish outcome fairness, and so more work is needed to balance process-based and outcome-based measures of fairness. The field of fair machine learning is simultaneously contentious and important, and we are excited to contribute towards the goal of constructing collectively fair decision-making processes without overly sacrificing efficiency.

References

[1] M. Altman, A. Wood, and E. Vayena, "A harm-reduction framework for algorithmic fairness," *IEEE Security & Privacy*, vol. 16, no. 3, pp. 34–45, 2018.

[2] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, *There's software used across the country to predict future criminals. and it's biased against blacks*, 2016. [Online]. Available: `www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing`.

[3] T. Araujo, N. Helberger, S. Kruikemeier, and C. H. De Vreese, "In AI we trust? Perceptions about automated decision-making by artificial intelligence," *AI & SOCIETY*, pp. 1–13, 2020.

[4] I. Ayres, "Outcome tests of racial disparities in police practices," *Justice Research and Policy*, vol. 4, no. 1-2, pp. 131–142, 2002.

[5] M. Baig, E. Zhang, R. Robinson, E. Ullah, and R. Whitakker, "Evaluation of Patients at Risk of Hospital Readmission (PARR) and LACE risk score for New Zealand context," *Studies in Health Technology and Informatics*, vol. 252, pp. 21–26, 2018.

[6] S. Barocas, "What is the problem to which fair machine learning is the solution," in *AI Now Experts Workshop on Bias and Inclusion*, 2017.

[7] S. Barocas and A. D. Selbst, "Big data's disparate impact," *California Law Review*, vol. 104, pp. 671–732, 2016.

[8] R. Berk, "Accuracy and fairness for juvenile justice risk assessments," *Journal of Empirical Legal Studies*, vol. 16, no. 1, pp. 175–194, 2019.

[9] R. Berk, H. Heidari, S. Jabbari, M. Kearns, and A. Roth, "Fairness in criminal justice risk assessments: The state of the art," *arXiv preprint arXiv:1703.09207*, 2017.

[10] J. Billings, I. Blunt, A. Steventon, T. Georghiou, G. Lewis, and M. Bardsley, "Development of a predictive model to identify inpatients at risk of re-admission within 30 days of discharge (PARR-30)," *BMJ open*, vol. 2, no. 4, 2012.

[11] R. Binns, M. Van Kleek, M. Veale, U. Lyngs, J. Zhao, and N. Shadbolt, "'It's reducing a human being to a percentage'; perceptions of justice in algorithmic decisions," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2018, pp. 1–14.

[12] A. Brown, A. Chouldechova, E. Putnam-Hornstein, A. Tobin, and R. Vaithianathan, "Toward algorithmic accountability in public services: A qualitative study of affected community perspectives on algorithmic decision-making in child welfare services," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–12.

[13] T. Calders, F. Kamiran, and M. Pechenizkiy, "Building classifiers with independency constraints," in *2009 IEEE International Conference on Data Mining Workshops*, 2009, pp. 13–18.

[14] Centers for Disease Control and Prevention (CDC), "Incidence of sickle cell trait — United States, 2010," *MMWR Morb Mortal Wkly Rep. 2014*, vol. 63, no. 49, pp. 1155–1158, Dec. 2014. [Online]. Available: `https://www.cdc.gov/mmwr/preview/mmwrhtml/mm6349a3.htm?s_cid=mm6349a3_w`.

[15] A. Chouldechova, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments," *Big Data*, vol. 5, no. 2, pp. 153–163, 2017.

[16] A. Chouldechova, D. Benavides-Prado, O. Fialko, and R. Vaithianathan, "A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions," in *Conference on Fairness, Accountability and Transparency*, 2018, pp. 134–148.

[17] S. Corbett-Davies and S. Goel, "The measure and mismeasure of fairness: A critical review of fair machine learning," *arXiv preprint arXiv:1808.00023*, 2018.

[18] S. Corbett-Davies, E. Pierson, A. Feller, and S. Goel, "A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear.," *Washington Post*, Oct. 2016. [Online]. Available: `https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/`.

[19] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq, "Algorithmic decision making and the cost of fairness," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017, pp. 797–806.

[20] K. Crenshaw, "Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics," *The University of Chicago Legal Forum*, pp. 139–169, 1989.

[21] T. Dare and E. Gambrill, *Ethical analysis: Predictive risk models at call screening for Allegheny County*, 2017. [Online]. Available: https://www.alleghenycountyanalytics.us/wp-content/uploads/2019/05/Ethical-Analysis-16-ACDHS-26_PredictiveRisk_Package_050119_FINAL-2.pdf.

[22] DeGraffenreid v. GENERAL MOTORS ASSEMBLY DIV., ETC., 413 F. Supp. 143, 1976.

[23] B. Delahaye, *Machine learning in banking: How banks approve automated loans*, Aug. 2018. [Online]. Available: https://www.neurochaintech.io/machine-learning-in-banking-loan-decisions/.

[24] W. Dieterich, C. Mendoza, and T. Brennan, "COMPAS risk scales: Demonstrating accuracy equity and predictive parity," *Northpointe Inc*, 2016.

[25] B. J. Dietvorst, J. P. Simmons, and C. Massey, "Algorithm aversion: People erroneously avoid algorithms after seeing them err.," *Journal of Experimental Psychology: General*, vol. 144, no. 1, pp. 114–126, 2015.

[26] J. Dodge, Q. V. Liao, Y. Zhang, R. K. Bellamy, and C. Dugan, "Explaining models: An empirical study of how explanations impact fairness judgment," in *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 2019, pp. 275–285.

[27] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 2012, pp. 214–226.

[28] Employment Equity, "Employment equity in federally regulated workplaces," Dec. 1995. [Online]. Available: https://www.canada.ca/en/employment-social-development/corporate/portfolio/labour/programs/employment-equity.html.

[29] Equal Credit Opportunity Act (ECOA), "Subchapter iv—equal credit opportunity," Oct. 1974. [Online]. Available: https://www.govinfo.gov/content/pkg/USCODE-2011-title15/html/USCODE-2011-title15-chap41-subchapIV.htm.

[30] Equal Employment Opportunity Commission, "Uniform guidelines on employee selection procedures," *Code of Federal Regulations*, vol. 29, no. 1607, 1978.

[31] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015, pp. 259–268.

[32] A. W. Flores, K. Bechtel, and C. T. Lowenkamp, "False positives, false negatives, and false analyses: A rejoinder to machine bias: There's software used across the country to predict future criminals. And it's biased against blacks," *Federal Probation*, vol. 80, no. 2, pp. 38–46, 2016.

[33] J. R. Foulds, R. Islam, K. N. Keya, and S. Pan, "An intersectional definition of fairness," in *2020 IEEE 36th International Conference on Data Engineering*, IEEE, 2020, pp. 1918–1921.

[34] S. A. Friedler, C. Scheidegger, and S. Venkatasubramanian, "On the (im) possibility of fairness," *arXiv preprint arXiv:1609.07236*, 2016.

[35] S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton, and D. Roth, "A comparative study of fairness-enhancing interventions in machine learning," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2019, pp. 329–338.

[36] General Data Protection Regulation (GDPR), "Right of access," *L 100000 May 2016*, pp. 1–88, Apr. 2016. [Online]. Available: `https://gdpr-info.eu/issues/right-of-access/`.

[37] J. Graham, J. Haidt, S. Koleva, M. Motyl, R. Iyer, S. P. Wojcik, and P. H. Ditto, "Moral foundations theory: The pragmatic validity of moral pluralism," in *Advances in Experimental Social Psychology*, vol. 47, 2013, pp. 55–130.

[38] N. Grgić-Hlača, E. M. Redmiles, K. P. Gummadi, and A. Weller, "Human perceptions of fairness in algorithmic decision making: A case study of criminal risk prediction," in *Proceedings of the 2018 World Wide Web Conference*, 2018, pp. 903–912.

[39] N. Grgić-Hlača, A. Weller, and E. M. Redmiles, "Dimensions of diversity in human perceptions of algorithmic fairness," *arXiv preprint arXiv:2005.00808*, 2020.

[40] N. Grgić-Hlača, M. B. Zafar, K. P. Gummadi, and A. Weller, "The case for process fairness in learning: Feature selection for fair decision making," in *NIPS Symposium on Machine Learning and the Law*, 2016.

[41] ——, "Beyond distributive fairness in algorithmic decision making: Feature selection for procedurally fair learning," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[42] Griggs v. Duke Power Co., 401 U.S. 424, 1971.

[43] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *Advances in Neural Information Processing Systems*, 2016, pp. 3315–3323.

[44] G. Harrison, J. Hanson, C. Jacinto, J. Ramirez, and B. Ur, "An empirical study on the perceived fairness of realistic, imperfect machine learning models," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020, pp. 392–402.

[45] U. Hébert-Johnson, M. P. Kim, O. Reingold, and G. N. Rothblum, "Calibration for the (computationally-identifiable) masses," *arXiv preprint arXiv:1711.08513*, 2017.

[46] J. E. Johndrow and K. Lum, "An algorithm for removing sensitive information: Application to race-independent recidivism prediction," *The Annals of Applied Statistics*, vol. 13, no. 1, pp. 189–220, 2019.

[47] K. D. Johnson, D. P. Foster, and R. A. Stine, "Impartial predictive modeling: Ensuring fairness in arbitrary models," *arXiv preprint arXiv:1608.00528*, 2016.

[48] M. Joseph, M. Kearns, J. Morgenstern, S. Neel, and A. Roth, "Rawlsian fairness for machine learning," *arXiv preprint arXiv:1610.09559*, 2016.

[49] M. Joseph, M. Kearns, J. H. Morgenstern, and A. Roth, "Fairness in learning: Classic and contextual bandits," *Advances in Neural Information Processing Systems*, vol. 29, pp. 325–333, 2016.

[50] N. Kaufmann, T. Schulze, and D. Veit, "More than fun and money. worker motivation in crowdsourcing-a study on mechanical turk.," in *Americas Conference on Information Systems*, Detroit, Michigan, USA, vol. 11, 2011, pp. 1–11.

[51] M. Kearns, S. Neel, A. Roth, and Z. S. Wu, "Preventing fairness gerrymandering: Auditing and learning for subgroup fairness," in *International Conference on Machine Learning*, PMLR, 2018, pp. 2564–2572.

[52] R. Kennedy, P. Waggoner, and M. Ward, "Trust in public policy algorithms," *Available at SSRN 3339475*, 2018.

[53] A. Khademi, S. Lee, D. Foley, and V. Honavar, "Fairness in algorithmic decision making: An excursion through the lens of causality," in *The World Wide Web Conference*, 2019, pp. 2907–2914.

[54] J. Kleinberg, J. Ludwig, S. Mullainathan, and A. Rambachan, "Algorithmic fairness," in *AEA Papers and Proceedings*, vol. 108, 2018, pp. 22–27.

[55] J. Kleinberg and S. Mullainathan, "Simplicity creates inequity: Implications for fairness, stereotypes, and interpretability," in *Proceedings of the 2019 ACM Conference on Economics and Computation*, 2019, pp. 807–808.

[56] J. Kleinberg, S. Mullainathan, and M. Raghavan, "Inherent trade-offs in the fair determination of risk scores," *arXiv preprint arXiv:1609.05807*, 2016.

[57] J. Konow, "Which is the fairest one of all? A positive analysis of justice theories," *Journal of economic literature*, vol. 41, no. 4, pp. 1188–1239, 2003.

[58] M. J. Kusner, J. Loftus, C. Russell, and R. Silva, "Counterfactual fairness," in *Advances in Neural Information Processing Systems*, 2017, pp. 4066–4076.

[59] M. K. Lee, "Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management," *Big Data & Society*, vol. 5, no. 1, pp. 1–16, 2018.

[60] Y. Liu, G. Radanovic, C. Dimitrakakis, D. Mandal, and D. C. Parkes, "Calibrated fairness in bandits," *arXiv preprint arXiv:1707.01875*, 2017.

[61] E. Loepp and J. T. Kelly, "Distinction without a difference? An assessment of MTurk worker types," *Research & Politics*, vol. 7, no. 1, pp. 1–8, 2020.

[62] J. M. Logg, J. A. Minson, and D. A. Moore, "Algorithm appreciation: People prefer algorithmic to human judgment," *Organizational Behavior and Human Decision Processes*, vol. 151, pp. 90–103, 2019.

[63] J. R. Macey and G. P. Miller, "The community reinvestment act: An economic analysis," *Virginia Law Review*, pp. 291–348, 1993.

[64] F. Marcinkowski, K. Kieslich, C. Starke, and M. Lünich, "Implications of AI (un-) fairness in higher education admissions: The effects of perceived AI (un-) fairness on exit, voice and organizational reputation," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020, pp. 122–130.

[65] L. Menand, "The changing meaning of affirmative action," *The New Yorker*, Jan. 2020. [Online]. Available: `https://www.newyorker.com/magazine/2020/01/20/have-we-outgrown-the-need-for-affirmative-action`.

[66] A. K. Menon and R. C. Williamson, "The cost of fairness in binary classification," in *Conference on Fairness, Accountability and Transparency*, 2018, pp. 107–118.

[67] A. Narayanan, "21 fairness definitions and their politics," in *Conference on Fairness, Accountability, and Transparency*, Feb. 2018.

[68] J. Niklas, K. Sztandar-Sztanderska, and K. Szymielewicz, "Profiling the unemployed in Poland: Social and political implications of algorithmic decision making," *Fundacja Panoptykon, Warsaw Google Scholar*, 2015.

[69] D. Pessach and E. Shmueli, "Algorithmic fairness," *arXiv preprint arXiv:2001.09784*, 2020.

[70] E. Pierson, "Demographics and discussion influence views on algorithmic fairness," *arXiv preprint arXiv:1712.09124*, 2017.

[71] G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, and K. Q. Weinberger, "On fairness and calibration," in *Advances in Neural Information Processing Systems*, 2017, pp. 5680–5689.

[72] B. Qureshi, F. Kamiran, A. Karim, S. Ruggieri, and D. Pedreschi, "Causal inference for social discrimination reasoning," *Journal of Intelligent Information Systems*, pp. 1–13, 2019.

[73] J. Rawls, *A Theory of Justice*. Harvard University Press, 2009.

[74] N. A. Saxena, K. Huang, E. DeFilippis, G. Radanovic, D. C. Parkes, and Y. Liu, "How do fairness definitions fare? Examining public attitudes towards algorithmic definitions of fairness," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 2019, pp. 99–106.

[75] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.

[76] C. Simoiu, S. Corbett-Davies, and S. Goel, "The problem of infra-marginality in outcome tests for discrimination," *The Annals of Applied Statistics*, vol. 11, no. 3, pp. 1193–1216, 2017.

[77] J. L. Skeem and C. T. Lowenkamp, "Risk, race, and recidivism: Predictive bias and disparate impact," *Criminology*, vol. 54, no. 4, pp. 680–712, 2016.

[78] T. Speicher, H. Heidari, N. Grgic-Hlaca, K. P. Gummadi, A. Singla, A. Weller, and M. B. Zafar, "A unified approach to quantifying algorithmic unfairness: Measuring individual & group unfairness via inequality indices," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018, pp. 2239–2248.

[79] M. Srivastava, H. Heidari, and A. Krause, "Mathematical notions vs. human perception of fairness: A descriptive approach to fairness for machine learning," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 2459–2468.

[80] S. Suri, D. G. Goldstein, and W. A. Mason, "Honesty in an online labor market," in *Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence*, 2011.

[81] W. J. Tastle and M. J. Wierman, "Consensus and dissention: A measure of ordinal dispersion," *International Journal of Approximate Reasoning*, vol. 45, no. 3, pp. 531–545, 2007.

[82] U.S. Census Bureau, "2014—2018 ACS 5-year data profile," 2018. [Online]. Available: `https://www.census.gov/acs/www/data/data-tables-and-tools/data-profiles/2018/`.

[83] B. Ustun, A. Spangher, and Y. Liu, "Actionable recourse in linear classification," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2019, pp. 10–19.

[84] R. Vaithianathan, E. Putnam-Hornstein, N. Jiang, P. Nand, and T. Maloney, "Developing predictive models to support child maltreatment hotline screening decisions: Allegheny County methodology and implementation," *Center for Social Data Analytics*, 2017.

[85] N. Van Berkel, J. Goncalves, D. Hettiachchi, S. Wijenayake, R. M. Kelly, and V. Kostakos, "Crowdsourcing perceptions of fair predictors for machine learning: A recidivism case study," *Proceedings of the ACM on Human-Computer Interaction - CSCW*, vol. 3, pp. 1–21, 2019.

[86] R. Wang, F. M. Harper, and H. Zhu, "Factors influencing perceived fairness in algorithmic decision-making: Algorithm outcomes, development procedures, and individual differences," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–14.

[87] A. Woodruff, S. E. Fox, S. Rousso-Schindler, and J. Warshaw, "A qualitative exploration of perceptions of algorithmic fairness," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018, pp. 1–14.

[88] M. B. Zafar, I. Valera, M. Gomez Rodriguez, and K. P. Gummadi, "Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment," in *Proceedings of the 26th International Conference on World Wide Web*, 2017, pp. 1171–1180.

[89] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations," in *International Conference on Machine Learning*, 2013, pp. 325–333.

[90] I. Žliobaitė, "On the relation between accuracy and fairness in binary classification," *arXiv preprint arXiv:1505.05723*, 2015.

Appendix A

Supplementary Tables and Figures

Table A.1 lists the consensus levels achieved in our initial survey as described in Section 4.1.1. For each feature in each domain, we provide the consensus (as measured by 1 minus the normalized Shannon entropy) that the feature exhibits each property; the mean consensus over properties for that feature; and the consensus about the fairness of the feature. The last two columns are correlated with a very strong Pearson correlation coefficient of 0.72.

Table A.2 shows the hypothetical sample sizes that would result in American census-representative [82] samples of each of our demographic groups per domain. In other words, because of our large total sample size, we would have equal total numbers of respondents from each demographic group as a American census-representative sample of the given size. Values listed as $< n$ are larger than our actual sample size per domain meaning that we over-sampled that demographic group.

Tables A.3 and A.4, as well as Figures A.1, A.2, A.3, and A.4 are identical to Table 4.1 and Figures 4.1 and 4.2 but using data without any Caucasians or males.

Domain	Feature	Relev.	C.O.	Reliab.	C.D.	C.G	C.C.	Vol.	Priv.	Prop. Avg.	Fairness
Bail	Current Charges	0.792	0.336	0.853	0.066	0.134	0.071	0.186	0.286	0.341	0.792
Bail	Criminal History	0.937	0.339	0.893	0.084	0.134	0.116	0.240	0.303	0.381	0.746
Bail	Substance Abuse	0.638	0.465	0.385	0.024	0.055	0.288	0.164	0.088	0.263	0.372
Bail	Employment & Living	0.379	0.238	0.427	0.121	0.067	0.171	0.379	0.238	0.253	0.231
Bail	Personality	0.153	0.313	0.254	0.084	0.098	0.048	0.124	0.037	0.139	0.106
Bail	Criminal Attitudes	0.175	0.159	0.153	0.010	0.006	0.056	0.308	0.059	0.116	0.132
Bail	Neighbourhood	0.195	0.036	0.374	0.198	0.091	0.086	0.201	0.250	0.179	0.217
Bail	Friends and Family	0.148	0.135	0.229	0.217	0.088	0.131	0.272	0.113	0.167	0.432
Bail	Social Life	0.124	0.126	0.241	0.043	0.031	0.063	0.349	0.185	0.145	0.329
Bail	Education	0.295	0.195	0.478	0.134	0.046	0.108	0.152	0.109	0.190	0.290
CPS	Child Demographics	0.149	0.093	0.292	0.145	0.200	0.147	0.156	0.039	0.153	0.198
CPS	Child CPS His.	0.615	0.393	0.549	0.102	0.074	0.027	0.113	0.144	0.252	0.749
CPS	Juvenile Justice His.	0.546	0.333	0.710	0.130	0.079	0.138	0.121	0.087	0.268	0.420
CPS	Child Welfare His.	0.457	0.159	0.378	0.041	0.063	0.076	0.154	0.116	0.180	0.355
CPS	Parent Demographics	0.205	0.057	0.394	0.295	0.210	0.114	0.209	0.067	0.194	0.174
CPS	Parent Welfare His.	0.232	0.161	0.395	0.164	0.091	0.080	0.106	0.102	0.167	0.365
CPS	Perp Demographics	0.218	0.043	0.311	0.279	0.161	0.181	0.095	0.066	0.169	0.138
CPS	Perp Welfare His.	0.285	0.221	0.443	0.092	0.090	0.006	0.121	0.072	0.166	0.320
CPS	All CPS His.	0.521	0.369	0.446	0.105	0.122	0.082	0.196	0.095	0.242	0.494
CPS	All Health His.	0.653	0.379	0.317	0.168	0.049	0.132	0.079	0.113	0.236	0.528
Hospital	Age	0.647	0.220	0.679	0.094	0.098	0.080	0.432	0.133	0.298	0.297
Hospital	Gender	0.338	0.088	0.584	0.122	0.029	0.116	0.057	0.098	0.179	0.107
Hospital	Race	0.096	0.115	0.264	0.205	0.088	0.036	0.528	0.120	0.181	0.226
Hospital	Place of Residence	0.110	0.087	0.324	0.117	0.048	0.049	0.490	0.222	0.181	0.090
Hospital	Hospital Treated	0.400	0.100	0.796	0.155	0.181	0.109	0.107	0.124	0.246	0.143
Hospital	Admission	0.714	0.091	0.856	0.175	0.171	0.106	0.109	0.248	0.309	0.487
Hospital	No. of Discharges	0.751	0.092	0.761	0.194	0.167	0.041	0.117	0.209	0.292	0.378
Hospital	History of Health	0.834	0.578	0.613	0.157	0.072	0.079	0.236	0.323	0.362	0.761
Insurance	Age	0.573	0.580	0.697	0.311	0.028	0.010	0.554	0.088	0.355	0.294
Insurance	Gender	0.117	0.080	0.419	0.292	0.119	0.022	0.141	0.023	0.152	0.355
Insurance	Marital and Family	0.060	0.096	0.256	0.021	0.020	0.053	0.326	0.037	0.109	0.195
Insurance	Employment	0.157	0.169	0.354	0.095	0.050	0.019	0.238	0.071	0.144	0.153
Insurance	Credit	0.231	0.128	0.378	0.070	0.032	0.224	0.141	0.099	0.163	0.198
Insurance	Education	0.122	0.058	0.256	0.144	0.020	0.021	0.255	0.015	0.111	0.265
Insurance	Place of Residence	0.095	0.095	0.378	0.145	0.092	0.022	0.341	0.068	0.155	0.164
Insurance	History of Health	0.359	0.493	0.124	0.229	0.161	0.200	0.113	0.420	0.262	0.202
Loan	Loan Amount	0.857	0.507	0.836	0.178	0.200	0.044	0.456	0.058	0.392	0.722
Loan	Income	0.858	0.717	0.573	0.065	0.107	0.024	0.271	0.078	0.337	0.896
Loan	Age	0.202	0.112	0.591	0.204	0.081	0.048	0.470	0.109	0.227	0.199
Loan	Gender	0.475	0.498	0.268	0.340	0.128	0.062	0.105	0.071	0.243	0.573
Loan	Marital and Family	0.133	0.019	0.390	0.061	0.033	0.020	0.351	0.024	0.129	0.084
Loan	No. of Dependents	0.194	0.070	0.456	0.054	0.044	0.010	0.074	0.037	0.118	0.094
Loan	Education	0.125	0.171	0.273	0.096	0.035	0.039	0.253	0.090	0.135	0.134
Loan	Employment	0.754	0.531	0.620	0.132	0.108	0.031	0.426	0.214	0.352	0.754
Loan	Credit	0.834	0.095	0.681	0.120	0.128	0.039	0.271	0.184	0.294	0.734
Loan	Property	0.679	0.074	0.433	0.055	0.077	0.017	0.190	0.127	0.207	0.392
Unem.	Age	0.177	0.314	0.520	0.354	0.171	0.208	0.324	0.081	0.269	0.196
Unem.	Gender	0.287	0.217	0.232	0.295	0.275	0.114	0.100	0.018	0.192	0.379
Unem.	Education	0.283	0.389	0.476	0.200	0.101	0.172	0.268	0.126	0.252	0.270
Unem.	Work History	0.570	0.379	0.659	0.141	0.124	0.120	0.191	0.155	0.292	0.440
Unem.	Skills	0.432	0.162	0.540	0.094	0.156	0.066	0.187	0.155	0.224	0.318
Unem.	Disability	0.335	0.495	0.256	0.350	0.167	0.193	0.258	0.153	0.276	0.164
Unem.	Time Unemployed	0.268	0.313	0.400	0.105	0.096	0.214	0.046	0.063	0.188	0.172
Unem.	Place of Residence	0.203	0.061	0.305	0.123	0.121	0.170	0.284	0.061	0.166	0.313
Unem.	Reason	0.125	0.021	0.180	0.125	0.065	0.028	0.383	0.032	0.120	0.039
Unem.	Initiative	0.263	0.304	0.141	0.212	0.073	0.053	0.433	0.015	0.187	0.126

Table A.1: Consensus levels achieved in our initial survey as described in Section 4.1.1. The values are 1 minus the Shannon entropy normalized between 0 and 1 (over responses bucketed into the categories "Unfair", "Neutral", and "Fair") so that 1 corresponds to complete consensus and 0 to complete disagreement. The eight columns corresponding to our eight properties are the levels of consensus reached when assigning that property to each feature, and the property average column is the average of the eight previous columns. The fairness column lists the levels of consensus achieved when rating how fair that feature is. The last two columns are correlated with a very strong Pearson correlation coefficient of 0.72.

Demographic Group	Bail	Unem.	CPS	Hos.	Loan	Ins.
18-29	>279	>282	>279	>277	>278	>274
30-39	>279	>282	>279	>277	>278	>274
40-49	>279	>282	>279	>277	>278	>274
50-59	195	203	219	227	195	219
60-69	114	170	91	68	68	136
70 up	63	109	47	47	16	16
Less than HS	0	8	0	16	8	24
High School	126	132	113	138	113	105
Associate or Diploma	>279	>282	>279	>277	>278	>274
Bachelor Degree	>279	>282	>279	>277	>278	>274
Graduate Degree	>279	>282	>279	240	>278	>274
Female	220	253	211	216	209	197
Male	>279	>282	>279	>277	>278	>274
<$25000	218	257	277	>277	238	208
$25000-$50000	>279	>282	>279	>277	>278	>274
$50000-$75000	>279	>282	>279	>277	>278	>274
$75000-$100000	248	240	>279	256	>278	>274
>$100000	100	82	82	72	100	111
Aboriginal	59	235	235	235	235	>274
Asian	>279	>282	>279	>277	>278	>274
Black	214	243	136	143	157	186
Caucasian	245	260	216	262	224	260
Hispanic	67	124	101	79	107	101

Table A.2: The hypothetical sample sizes for which our total number of respondents in each demographic group would be exactly representative of the United States population [82] rounded to the nearest whole number. Values expressed as >n are larger than the sample size that we actually had for that domain, indicating that we over-sampled that demographic group. For most of the demographic groups that we under-sampled, we still have more total respondents than an American representative sample of 100 people would have.

Domain	Within	Cross Trained	Cross Tested
Bail	$79.2 \pm 0.00098\%$	$79.0 \pm 0.018\%$	$77.0 \pm 0.018\%$
CPS	$83.0 \pm 0.00075\%$	$73.2 \pm 0.037\%$	$78.1 \pm 0.011\%$
Hospital	$77.0 \pm 0.00128\%$	$76.8 \pm 0.014\%$	$75.5 \pm 0.016\%$
Insurance	$78.0 \pm 0.00099\%$	$78.0 \pm 0.014\%$	$74.8 \pm 0.040\%$
Loan	$83.3 \pm 0.00101\%$	$77.0 \pm 0.018\%$	$82.2 \pm 0.011\%$
Unemployment	$77.7 \pm 0.00091\%$	$80.8 \pm 0.011\%$	$77.2 \pm 0.009\%$
All (Pooled)	$80.7 \pm 0.00037\%$	N/A	N/A

Table A.3: Identical to Table 4.1, but using data without Caucasians. Table 4.1 caption: The average accuracies and standard errors of the mean for the within-domain, all-domain, and cross-domain predictors (predicting fairness judgements using property assignments). *Within:* Accuracies obtained by training and testing the model within the listed domain (or over all pooled domains), each average is over 1000 50%/50% train/test splits. *Cross Trained:* Accuracies obtained by training in the listed domain and testing in each of the other domains individually, average is over the five other domains used to test. *Cross Tested:* Accuracies obtained by training in each of the other domains individually and testing in the listed domain, average is over the five other domains used to train.

Domain	Within	Cross Trained	Cross Tested
Bail	$81.8 \pm 0.00027\%$	$80.3 \pm 0.013\%$	$82.6 \pm 0.004\%$
CPS	$79.9 \pm 0.00075\%$	$79.6 \pm 0.023\%$	$78.1 \pm 0.015\%$
Hospital	$76.7 \pm 0.00085\%$	$80.7 \pm 0.008\%$	$75.3 \pm 0.010\%$
Insurance	$80.2 \pm 0.00088\%$	$78.3 \pm 0.027\%$	$78.7 \pm 0.013\%$
Loan	$87.0 \pm 0.00058\%$	$79.3 \pm 0.010\%$	$85.3 \pm 0.007\%$
Unemployment	$78.8 \pm 0.00069\%$	$81.1 \pm 0.019\%$	$79.3 \pm 0.004\%$
All (Pooled)	$81.4 \pm 0.00027\%$	N/A	N/A

Table A.4: Identical to Table 4.1, but using data without males. Table 4.1 caption: The average accuracies and standard errors of the mean for the within-domain, all-domain, and cross-domain predictors (predicting fairness judgements using property assignments). *Within:* Accuracies obtained by training and testing the model within the listed domain (or over all pooled domains), each average is over 1000 50%/50% train/test splits. *Cross Trained:* Accuracies obtained by training in the listed domain and testing in each of the other domains individually, average is over the five other domains used to test. *Cross Tested:* Accuracies obtained by training in each of the other domains individually and testing in the listed domain, average is over the five other domains used to train.

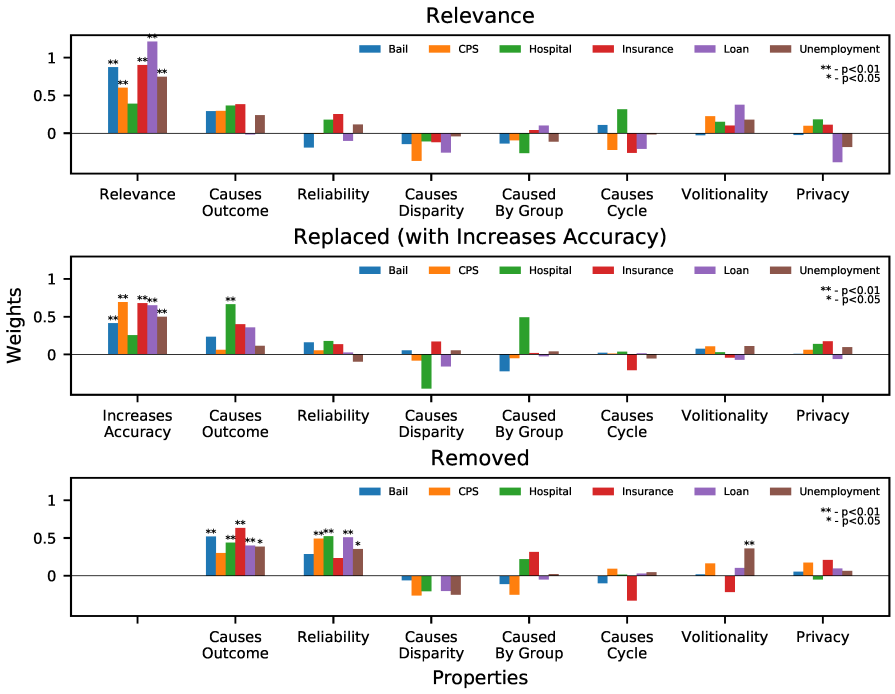

Figure A.1: Identical to Figure 4.1, but using data without Caucasians. Figure 4.1 caption: The weights associated with each property in every domain for the full property models of the initial survey (Relevance), the survey that replaced *relevance* with *increases accuracy* (Replaced), and the survey that removed *relevance* all together (Removed). Significant weights (using a linear regression t-test with the null hypothesis that the coefficient values are equal to zero) at the $p < 0.05$ (*) and $p < 0.01$ (**) levels are indicated.

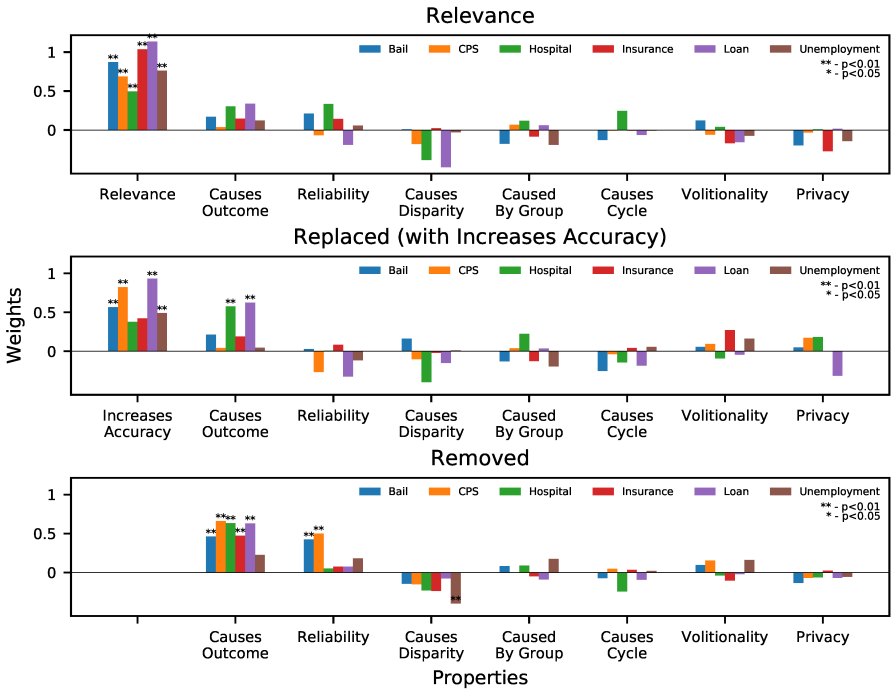

Figure A.2: Identical to Figure 4.1, but using data without males. Figure 4.1 caption: The weights associated with each property in every domain for the full property models of the initial survey (Relevance), the survey that replaced *relevance* with *increases accuracy* (Replaced), and the survey that removed *relevance* all together (Removed). Significant weights (using a linear regression t-test with the null hypothesis that the coefficient values are equal to zero) at the $p < 0.05$ (*) and $p < 0.01$ (**) levels are indicated.

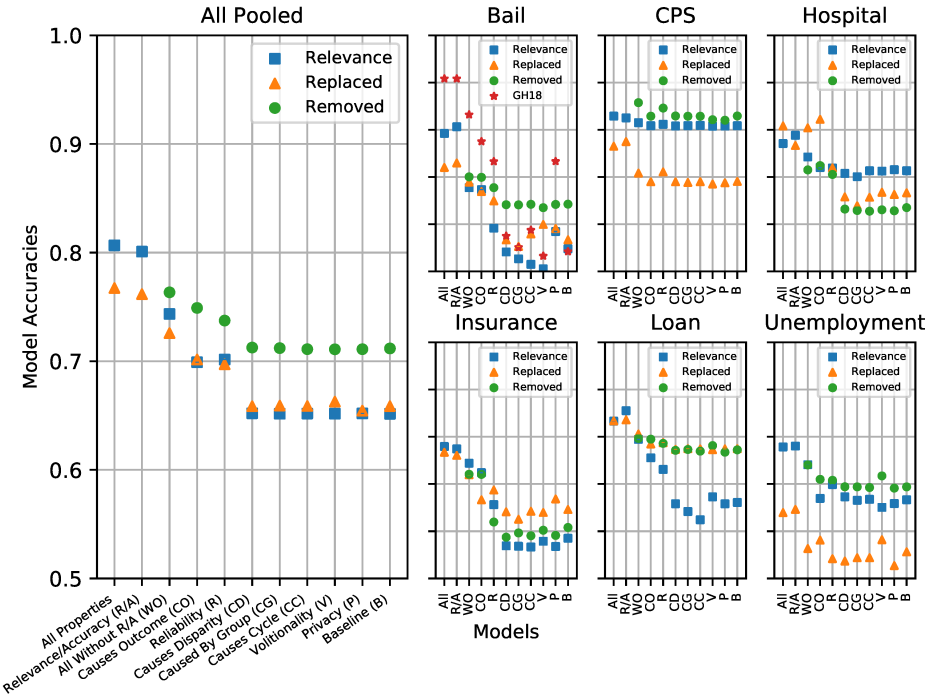

Figure A.3: Identical to Figure 4.2, but using data without Caucasians. Figure 4.2 caption: The accuracy levels achieved per model. For each domain (and all pooled), the accuracies achieved from the initial survey (Relevance) and additional surveys (Replaced and Removed) are shown for the following models where applicable: using all eight properties, using only *relevance/increases accuracy*, using all seven properties excluding *relevance/increases accuracy*, and each of the other seven properties individually. Additionally for comparison, we show the baseline accuracies that are achieved by guessing only either "Fair" or "Unfair" (whichever does better). Error bars are present for each point but are too small to be visible.

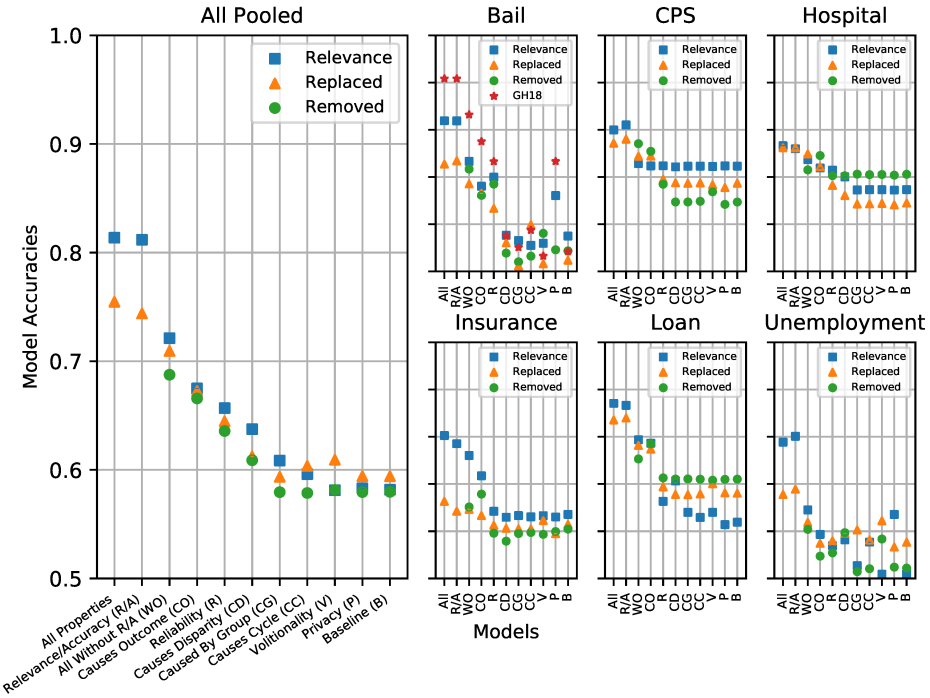

Figure A.4: Identical to Figure 4.2, but using data without males. Figure 4.2 caption: The accuracy levels achieved per model. For each domain (and all pooled), the accuracies achieved from the initial survey (Relevance) and additional surveys (Replaced and Removed) are shown for the following models where applicable: using all eight properties, using only *relevance/increases accuracy*, using all seven properties excluding *relevance/increases accuracy*, and each of the other seven properties individually. Additionally for comparison, we show the baseline accuracies that are achieved by guessing only either "Fair" or "Unfair" (whichever does better). Error bars are present for each point but are too small to be visible.

Appendix B

Survey Questions

B.1 Fairness Judgement Questions

For all surveys, participants were asked to judge how fair it is for specific features to be used by ADMs in certain domains. Listed below are the two possible descriptions that were given for each domain, followed by the exact fairness questions asked in that domain, which were answered on a 7-point Likert scale ("Very Unfair", "Unfair", "Somewhat Unfair", "Neutral", "Somewhat Fair", "Fair", or "Very Fair"). In our first pilot survey (Section 3.3), respondents were randomly assigned to receive either the first (mentions machine learning) or second (does not mention machine learning) descriptions, and then asked the 16 starred (*) questions. For all other surveys, respondents were randomly assigned one domain and asked all questions for that domain in a random order. For the full surveys as HTML files, the data files, and our code, see https://github.com/michelealbach/cross-domain-fairness.

B.1.1 Bail

Version 1: "When judges are deciding whether or not to grant bail to a defendant in the United States, they must consider the chance that the defendant will reoffend if set free. COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) is a machine learning computer program used in some US courts to predict the risk that a defendant will reoffend. The program uses information that is obtained from a questionnaire filled out by the defendant and outputs a risk score. Judges can then use the risk score to

help with bail decisions. The following questions are about how fair it would be, in your opinion, if the program used certain information to calculate the score. For each question, please choose an option from very unfair to very fair."

Version 2: "When judges are deciding whether or not to grant bail to a defendant in the United States, they must consider the chance that the defendant will reoffend if set free. In order to predict if they will reoffend, some judges use information that is obtained from a questionnaire filled out by the defendant. The following questions are about how fair it would be, in your opinion, if judges used certain information to make these decisions. For each question, please choose an option from very unfair to very fair."

- How fair is it to determine if a person can be released on bail using information about the **current charges**?
- *How fair is it to determine if a person can be released on bail using information about their **criminal history**?
- *How fair is it to determine if a person can be released on bail using information about their **substance abuse history**?
- How fair is it to determine if a person can be released on bail using information about their **stability of employment and living situation**?
- How fair is it to determine if a person can be released on bail using information about their **personality** (for example, obtained through the question "do you have the ability to "sweet talk" people into getting what you want?")?
- How fair is it to determine if a person can be released on bail using information about their **criminal attitudes** (for example, obtained through the question "do you think that a hungry person has a right to steal?")?
- *How fair is it to determine if a person can be released on bail using information about the **safety of their neighbourhood**?

- How fair is it to determine if a person can be released on bail using information about the **criminal history of their friends and family**?
- How fair is it to determine if a person can be released on bail using information about the **quality of their social life and free time** (for example, obtained through the question "do you often feel left out of things?")?
- *How fair is it to determine if a person can be released on bail using information about their **level of education**?

B.1.2 CPS

Version 1: "In the county of Allegheny, Pennsylvania, USA, a machine learning computer program called the Allegheny Family Screening Tool is used by the county's CPS (Child Protective Services) to help hotline staff decide whether or not a tip should be screened in, meaning to start a CPS investigation. The program uses information about the family/people involved in the tip and outputs a risk score that predicts the probability that a child will be removed from the home within 2 years if the tip is screened in. The following questions are about how fair it would be, in your opinion, if the program used certain information to calculate the score. For each question, please choose an option from very unfair to very fair."

Version 2: "In the county of Allegheny, Pennsylvania, USA, hotline staff at the county's CPS (Child Protective Services) receive tips and must decide whether or not they should be screened in, meaning to start a CPS investigation. The staff members use information about the family/people involved in the tip to predict the chance that a child would be removed from the home if the tip is screened in. The following questions are about how fair it would be, in your opinion, if the staff used certain information to make these decisions. For each question, please choose an option from very unfair to very fair."

- How fair is it to determine whether or not a tip should be screened in to CPS using information about the **demographics of the child victim** (excluding race)?

- How fair is it to determine whether or not a tip should be screened in to CPS using information about the **CPS history of the child victim**?
- How fair is it to determine whether or not a tip should be screened in to CPS using information about the **juvenile justice history of the child victim**?
- How fair is it to determine whether or not a tip should be screened in to CPS using information about the **public welfare history of the child victim**?
- How fair is it to determine whether or not a tip should be screened in to CPS using information about the **demographics of the parents or other involved adults** (excluding race)?
- How fair is it to determine whether or not a tip should be screened in to CPS using information about the **public welfare history of the parents or other involved adults**?
- How fair is it to determine whether or not a tip should be screened in to CPS using information about the **demographics of the alleged perpetrators** (excluding race)?
- How fair is it to determine whether or not a tip should be screened in to CPS using information about the **public welfare history of the alleged perpetrators**?
- How fair is it to determine whether or not a tip should be screened in to CPS using information about the **CPS history of all individuals named in the referral**?
- How fair is it to determine whether or not a tip should be screened in to CPS using information about the **behavioural health history of all individuals named in the referral**?

B.1.3 Hospital Resources

Version 1: "When a patient is released from the hospital, it is beneficial for doctors to be aware of the chance that the person will be readmitted in the near future so that they can provide additional care to prevent readmission. PARR-30 (Patients at Risk of Readmission within 30 days) is a machine learning computer program that uses information about the patient and outputs a risk score representing the chance that they will be readmitted within 30 days. The following questions are about how fair it would be, in your opinion, if the program used certain information to calculate the score. For each question, please choose an option from very unfair to very fair."

Version 2: "When a patient is released from the hospital, it is beneficial for doctors to be aware of the chance that the person will be readmitted in the near future so that they can provide additional care to prevent readmission. Doctors estimate the chance that a patient will be readmitted using information about the patient. The following questions are about how fair it would be, in your opinion, if the doctors used certain information to decide whether to provide additional care in order to prevent readmission. For each question, please choose an option from very unfair to very fair."

- *How fair is it to allocate additional doctor care to prevent readmission using information about a patient's **age**?
- *How fair is it to allocate additional doctor care to prevent readmission using information about a patient's **gender**?
- How fair is it to allocate additional doctor care to prevent readmission using information about a patient's **race**?
- *How fair is it to allocate additional doctor care to prevent readmission using information about a patient's **place of residence**?
- How fair is it to allocate additional doctor care to prevent readmission using information about the **hospital where they were treated**?

- How fair is it to allocate additional doctor care to prevent readmission using information about a patient's **current hospital admission**?

- How fair is it to allocate additional doctor care to prevent readmission using information about a patient's **number of emergency hospital discharges**?

- *How fair is it to allocate additional doctor care to prevent readmission using information about a patient's **history of major health conditions**?

B.1.4 Insurance

Version 1: "When deciding on an applicant's insurance rates, insurance companies often use machine learning computer programs to predict the levels of risk associated with each applicant. These programs use information about the applicant and their car or home to return a score which is then used to decide what rates to set. The following questions are about how fair it would be, in your opinion, if the program used certain information to calculate the score. For each question, please choose an option from very unfair to very fair, and then explain your judgement by checking each box that applies."

Version 2: "When deciding on an applicant's insurance rates, insurance companies must predict the levels of risk associated with each applicant. To do so, employees use information about the applicant and their car or home to calculate a score which is then used to decide what rates to set. The following questions are about how fair it would be, in your opinion, if the employees used certain information to calculate the score. For each question, please choose an option from very unfair to very fair, and then explain your judgement by checking each box that applies."

- How fair is it to determine an applicant's insurance rates using information about their **age**?

- How fair is it to determine an applicant's insurance rates using information about their **gender**?

- How fair is it to determine an applicant's insurance rates using information about their **marital and family status**?
- How fair is it to determine an applicant's insurance rates using information about their **employment status**?
- How fair is it to determine an applicant's insurance rates using information about their **credit history**?
- How fair is it to determine an applicant's insurance rates using information about their **level of education**?
- How fair is it to determine an applicant's insurance rates using information about their **place of residence**?
- How fair is it to determine an applicant's insurance rates using information about their **history of major health conditions**?

B.1.5 Loan

Version 1: "When deciding whether or not to approve a loan application, banks often use machine learning computer programs to predict the chance that the person applying will default and be unable to pay back the bank. These programs use information about the applicant and return a score which is then used to decide whether to approve a loan. The following questions are about how fair it would be, in your opinion, if the program used certain information to calculate the score. For each question, please choose an option from very unfair to very fair."

Version 2: "When deciding whether or not to approve a loan application, banks need to consider the chance that the person applying will default and be unable to pay back the bank. To do so, employees use information about the applicant to calculate a score which is then used to decide whether to approve a loan. The following questions are about how fair it would be, in your opinion, if the employees used certain information to calculate the score. For each question, please choose an option from very unfair to very fair."

- How fair is it to determine an applicant's eligibility for a loan using information about the **loan amount**?
- *How fair is it to determine an applicant's eligibility for a loan using information about their **income**?
- *How fair is it to determine an applicant's eligibility for a loan using information about their **age**?
- How fair is it to determine an applicant's eligibility for a loan using information about their **gender**?
- *How fair is it to determine an applicant's eligibility for a loan using information about their **marital and family status**?
- How fair is it to determine an applicant's eligibility for a loan using information about their **number of dependents**?
- How fair is it to determine an applicant's eligibility for a loan using information about their **level of education**?
- How fair is it to determine an applicant's eligibility for a loan using information about their **employment status**?
- *How fair is it to determine an applicant's eligibility for a loan using information about their **credit history**?
- How fair is it to determine an applicant's eligibility for a loan using information about their **owned property value**?

B.1.6 Unemployment

Version 1: "In 2014, the Polish government introduced a machine learning computer program that is used to help with decision making for unemployment benefits. When an unemployed person asks for aid, the program uses information about them obtained from a questionnaire and outputs a score for their employment potential. Their score is used to determine which financial

aid benefits they are eligible for. The following questions are about how fair it would be, in your opinion, if the program used certain information to calculate the score. For each question, please choose an option from very unfair to very fair."

Version 2: "In 2014, the Polish government introduced a system that is used to help with decision making for unemployment benefits. When an unemployed person asks for aid, an employee asks them a series of questions and uses the answers to calculate a score for their employment potential. Their score is used to determine which financial aid benefits they are eligible for. The following questions are about how fair it would be, in your opinion, if employees used certain information to calculate the score. For each question, please choose an option from very unfair to very fair."

- *How fair is it to determine a person's eligibility for unemployment aid using their **age**?
- *How fair is it to determine a person's eligibility for unemployment aid using their **gender**?
- *How fair is it to determine a person's eligibility for unemployment aid using their **level of education**?
- How fair is it to determine a person's eligibility for unemployment aid using their **work history over the last 5 years**?
- How fair is it to determine a person's eligibility for unemployment aid using their **professional skills**?
- *How fair is it to determine a person's eligibility for unemployment aid using their **degree of disability**?
- How fair is it to determine a person's eligibility for unemployment aid using their **time spent unemployed**?
- How fair is it to determine a person's eligibility for unemployment aid using their **place of residence**?

- How fair is it to determine a person's eligibility for unemployment aid using their **reason for wanting a job** (other than income)?
- How fair is it to determine a person's eligibility for unemployment aid using their **initiative** (for example, obtained through the question "what are you able to do to increase your chances of finding a job?")?

B.2 Necessity and Sufficiency Explanatory Questions

In our second pilot study (Section 3.4.1), participants were asked to explain their fairness judgements after every question by checking any number of the properties or by filling in a blank text box. Depending on if they had selected a "Fair" (or "Neutral") option or an "Unfair" option, participants saw one of the following question versions after every feature fairness judgement question. [Feature] is replaced with the bolded part of the previous question.

Version 1 (Answered "Fair" or "Neutral"): "Please explain why you think it is fair / neither fair nor unfair to use information about their [feature]. You may do so by checking any number of the following suggestions or by filling in the blank option at the bottom."

- Their [feature] can be assessed reliably
- Their [feature] is relevant to [outcome] decisions
- Their [feature] is not private
- Their [feature] is caused by choices made by the person
- Their [feature] can cause them to have a high risk level
- Making this decision using information about their [feature] cannot cause a vicious cycle
- Making this decision using information about their [feature] cannot have negative effects on certain groups of people that are protected by law (eg., based on race, gender, age, religion, national origin, disability status)

- Their [feature] cannot be caused by their belonging to a group protected by law (eg., based on race, gender, age, religion, national origin, disability status)
- Other: (fill in the blank)

Version 2 (Answered "Unfair"): "Please explain why you think it is unfair to use information about their [feature]. You may do so by checking any number of the following suggestions or by filling in the blank option at the bottom."

- Their [feature] cannot be assessed reliably
- Their [feature] is not relevant to [outcome] decisions
- Their [feature] is private
- Their [feature] is not caused by choices made by the person
- Their [feature] cannot cause them to have a high risk level
- Making this decision using information about their [feature] can cause a vicious cycle
- Making this decision using information about their [feature] can have negative effects on certain groups of people that are protected by law (eg., based on race, gender, age, religion, national origin, disability status)
- Their [feature] can be caused by their belonging to a group protected by law (eg., based on race, gender, age, religion, national origin, disability status)
- Other: (fill in the blank)

B.3 Property Assignment Questions

In all of our non-pilot surveys, participants were asked to rate how strongly they felt the features held each of the eight properties (plus the ninth property

increases accuracy in our later survey), randomly before or after making their fairness judgements. Listed below are the property assigning statements that participants were asked to rate how much they agreed with on a 7-point Likert scale ("Strongly Disagree", "Disagree", "Somewhat Disagree", "Neutral", "Somewhat Agree", "Agree", or "Strongly Agree") for each feature.

- Information about their [feature] can be assessed **reliably**.
- Information about their [feature] is **relevant** to [outcome] decisions. / Using information about their [feature] would increase the **accuracy** of [outcome] decisions.
- Information about their [feature] is **private**.
- A person can change their [feature] by making a **choice or decision**.
- Their [feature] can **cause** them to [have an outcome].
- Making this decision using information about their [feature] can cause a **vicious cycle**.
- Making this decision using information about their [feature] can have **negative effects on certain groups** of people that are protected by law (e.g., based on race, gender, age, religion, national origin, disability status).
- Their [feature] can be **caused by their belonging to a group** protected by law (e.g., race, gender, age, religion, national origin, disability status).

B.4 Demographic Questions

With the exception of the first pilot survey, all participants were asked the following (optional) demographic questions.

1. Select the box that corresponds to your age:
    - 18-29

- 30-39
- 40-49
- 50-59
- 60-69
- 70 or above

2. Select the box that describes your current completed level of education:
    - Less than high school degree
    - High school degree or equivalent
    - Associate degree or diploma
    - Bachelor degree
    - Graduate degree
3. Select the box that best describes how you identify:
    - Female
    - Male
    - Nonbinary
    - Other: (fill in the blank)
4. Select the box that best describes your annual household income in American dollars:
    - Less than $25,000
    - $25,000 to $50,000
    - $50,000 to $75,000
    - $75,000 to $100,000
    - Over $100,000
5. Select all boxes that apply to you:
    - Aboriginal/Indigenous

- Asian
- Black/African
- Caucasian
- Hispanic/Latinx
- Other: (fill in the blank)