Limitations of the "Four-Fifths Rule" and Statistical Parity Tests for Measuring Fairness*

Manish Raghavan
Massachusetts Institute of Technology
mragh@mit.edu

Pauline Kim
Washington University School of Law
kim@wustl.edu

Abstract

To ensure the fairness of algorithmic decision systems, such as employment selection tools, computer scientists and practitioners often refer to the so-called "four-fifths rule" as a measure of a tool's compliance with anti-discrimination law. This reliance is problematic because the "rule" is in fact not a legal rule for establishing discrimination, and it offers a crude test that will often be over- and under-inclusive in identifying practices that warrant further scrutiny. The "four-fifths rule" is one of a broader class of statistical tests, which we call Statistical Parity Tests (SPTs), that compare selection rates across demographic groups. While some SPTs are more statistically robust, all share some critical limitations in identifying disparate impacts retrospectively. When these tests are used prospectively as an optimization objective shaping model development, additional concerns arise about the development process, behavioral incentives, and gameability. In this article, we discuss the appropriate role for SPTs in algorithmic governance. We suggest a combination of measures that take advantage of the additional information present during prospective optimization, providing greater insight into fairness considerations when building and auditing models.

1 Introduction

Algorithmic tools have become increasingly common in a variety of social domains like consumer finance, housing, employment, and criminal law enforcement. For example, in the employment context, a typical use case involves algorithms that screen job applicants to determine which candidates should be advanced in the hiring process. The applicant provides information, such as a resume, responses to a questionnaire, or even a recorded video, which is then broken down to discrete data points which are analyzed by the algorithm to make a recommendation [7, 41]. These algorithms typically entail models trained on historical data, and are often developed by third-party firms specializing in algorithmic assessments.

As awareness has grown that algorithms can discriminate, computer scientists and practitioners have sought to develop methods to ensure that models are fair. In doing so, many such efforts reference the so-called "four-fifths rule" as a measure of a tool's compliance with anti-discrimination law. The "four-fifths rule" examines the ratio of selection rates across relevant demographic characteristics. For a given selection tool or practice, it asks whether the selection rate of a disadvantaged group is less than four-fifths, or 80%, of the selection rate of an advantaged group. So, for example, if 67% of black applicants and 90% of white applicants are selected for a benefit, the ratio of selection is 67/90 or 74%. Because the ratio is less than 4/5 or 80%, the practice would be judged to have a disparate impact on black applicants.

*A full version of this paper can be found at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4624571

Workshop on Regulatable Machine Learning at the 37th Conference on Neural Information Processing Systems (RegML @ NeurIPS 2023).

The focus on the four-fifths ratio has its origins in law—specifically in employment discrimination law—but its use as a metric for measuring fairness in algorithms is problematic for two primary reasons. First, it is not a legal rule and has never been. To the extent that developers have turned to a four-fifths ratio as a way of ensuring compliance with anti-discrimination law, they are mistaken. It does not provide the legal definition of discrimination, and courts have generally rejected it as a determinative test for finding a prima facie case of disparate impact. Second, putting aside the inaccurate understanding of the law, a "four-fifths rule" is a poor measure of discrimination because it is a crude statistical measure that will often be over- and under-inclusive in identifying practices that warrant further scrutiny.

The "four-fifths rule" is one of a broader class of statistical tests that we call Statistical Parity Tests, or SPTs. SPTs seek to measure fairness by comparing positive outcomes across groups—for example, the rate at which black and white candidates are selected for a job. Examples of SPTs include Fisher's exact test and the chi-squared test [39]. Some of the deficiencies of a four-fifths rule are addressed by SPTs that are more statistically robust; nevertheless, all of these tests share some critical limitations.

In the legal context, courts look to statistical tests to determine whether a prima facie case exists that a particular practice has a disparate impact, warranting further scrutiny. In that sense, the inquiry is retrospective—it asks whether a challenged practice or test has caused disproportionate disadvantage for marginalized groups. When developers use the 4/5 ratio as a fairness metric, however, they rely on it *prospectively* as an optimization objective shaping model development. This new context raises additional concerns about the development process, behavioral incentives, and gameability.

On the other hand, prospective testing that goes beyond a simplistic four-fifths rule can provide both firms and regulators with more comprehensive insights into the properties of a model than are available after the fact. This potential for greater insight results from differences in the availability of information. In a retrospective analysis, a regulator has access to limited information, preventing them from using some sophisticated measures of discrimination, whereas a firm or auditor conducting ex ante testing during the model development and validation phases has more information in some ways. Pre-deployment audits allow regulators to take advantage of the information available to model developers ex ante, and in what follows, we argue that regulators can and should leverage this additional information to incentivize the development of fair algorithms.

In this Essay, we analyze the use of the four-fifths rule of thumb (and more statistically robust variants) for algorithmic employment decisions, primarily focusing on employment selection tools as an illustrative use case. In Section 2, we analyze the limitations of the "four-fifths rule" and other SPTs as applied to algorithms. We argue that in isolation they are too blunt as instruments for detecting whether a practice caused discrimination, and when used prospectively to optimize algorithmic hiring assessments, they can create further problems. In Section 3, we discuss the appropriate role of SPTs in algorithmic governance, noting some positive aspects of SPTs and suggesting ways to combine them with other types of prospective testing to provide comprehensive insights into the fairness properties of a model. For additional background on the "four-fifths rule" in both law and computer science, see Appendices A and B respectively.

2 Limitations of a "four-fifths rule" and SPTs generally

The four-fifths ratio has become a common metric for evaluating model fairness, even though it does not actually reflect the legal test for disparate impact discrimination—but does that discrepancy matter? One might argue that regardless of its legal status, it offers a useful metric for measuring discriminatory effects that should guide model development. However, relying on the four-fifths ratio as a fairness metric is problematic for a number of reasons.

First, on its own, it is a poor test for determining when a selection algorithm warrants closer legal scrutiny. There is a further risk that if compliance with a four-fifths ratio comes to be seen as an industry "best practice," courts lacking expertise in statistics and machine learning may rely on it as a legal rule despite its significant limitations. Second, to the extent that a four-fifths rule or other SPT is used to guide the model development process, it can distort incentives in ways that prioritize formal compliance without actually addressing model unfairness. Here, we detail the limits of SPTs for both retrospective and prospective analyses.

2.1 The prima facie case: a retrospective view

When statistical parity tests are used to determine whether a prima facie case exists, they are retrospective in orientation. They take data about actual applicants, examine the outputs of the model, and determine whether it has disproportionately screened out disadvantaged or marginalized groups. When used for this type of retrospective examination, SPTs may be a poor tool for deciding which cases warrant further legal scrutiny, as they can be both over- and under-inclusive. For example, some practices that produce ratios less than 4/5 may not constitute discrimination; other practices with ratios above 4/5 may still warrant close scrutiny. Even when statistical significance is taken into account, SPTs have limitations when used to establish a prima facie case due to several factors: measuring statistical significance is limited by the size of the dataset; the conclusions of SPTs are heavily dependent upon the selectivity of the underlying practice; and SPTs are not easily applied to practices that do not produce binary outcomes.

Statistical significance. On its own, the 4/5 ratio only measures effect size, not statistical significance. For small sample sizes, the 4/5 ratio is thus quite unreliable; the exclusion or inclusion of a single data point can easily alter the conclusions of this test. Recognizing this, courts have relied on other SPTs that provide information on both the magnitude of an effect (how different the selection rates are) and its statistical significance (the likelihood of such an effect occurring due to chance).

While more robust SPTs overcome some of the limitations of relying on the 4/5 ratio, reporting statistical significance is not a panacea. Statistical significance tests whether there is sufficient evidence to reject a "null hypothesis" that a practice is not discriminatory. Even if a practice is discriminatory (meaning the "null hypothesis" is false), a dataset may simply be too small for an SPT to draw a statistically significant conclusion. And on the other hand, given enough data, an SPT is likely to find statistically significant effects simply because with large amounts of data from the real world, a test can detect even small, idiosyncratic biases that may not warrant legal action [27]. In effect, a statistically significant result from an SPT can be as much indicative of sufficiently large samples as it is of discriminatory behavior. In practice, courts do not use a fixed rule when establishing a prima facie case; instead, they take both effect size and statistical significance into account when making judgements about whether practices warrant deeper scrutiny, and such an approach would also be appropriate if the challenged practice was an algorithmic system.

Selectivity. The effect of applying a 4/5 rule also depends upon the selectivity of an employer's process [39]. Examining the *ratio* of selection rates means that more selective processes will more likely result in a finding of disparate impact. Compare, for example, a highly selective process in which 2% of white applicants and 1.5% of black applicants are hired with a less selective process in which 90% of white and 73% of black applicants are hired. The first scenario, where the difference in selection rates is 0.5%, falls below the four-fifths ratio because the ratio of selection rates is 0.75. In the second scenario, despite the larger absolute difference in selection rates of 17%, the practice does not violate a four-fifths rule. Without knowing more about the specific selection procedure and the type of job at issue, it is difficult to assess whether these judgments are correct. However, it is not at all obvious that the more selective procedure, which affects far fewer black applicants, poses the more serious threat to equal opportunity.

Some SPTs—for example, the z-test—detect *differences* in selection rates instead of analyzing the ratio of selection rates. Again, the selectivity of the procedure influences whether a test will produce statistically significant results. In particular, differences in selection rates are easier (i.e., require less data) to detect when selection rates are extremely low or high—e.g., close to 0 or 1.[2] Thus, the sensitivity of an SPT to differences in selection rates depends on the selectivity of the practice, regardless of whether the test considers the ratios differences.

Beyond binary outcomes. A four-fifths rule is easiest to apply where an algorithm sorts candidates into or out of a pool—a binary decision. Algorithmic prediction tools, however, generally work in continuous scores. For example, they may predict the likelihood of particular outcomes—e.g. whether this individual will be a successful employee—not fixed certainties. It is up to the humans who design

[2]This is because distinguishing between distributions is harder when they have high variance, and variance is maximized when selection rates are close to 0.5. For example, detecting differences between selection rates of 0.001 and 0.051 requires less data than distinguishing between 0.5 and 0.55.

or implement these tools to decide what cut-off score to use to make the selection decision. If the distribution of scores looks different for different groups, then the relative selection rates of different groups will vary depending upon which cut-off score is chosen. Under a legal regime narrowly focused on the 4/5 ratio, the cut-off might be selected with an eye to increasing the selection ratio so that it falls above 4/5, even though the underlying rankings significantly favor one group over another. Very often, algorithms are used as screening tools, with subsequent decisions further narrowing the pipeline of candidates. If rankings influence these later decisions, the fact that an early screening tool complies with a 4/5 rule may obscure from view the downstream impacts of unequal predictions.

A new and growing class of AI techniques does not directly produce numerical estimates of candidate quality, but instead, seeks to infer relevant information for use in making judgments downstream. For example, commercially available resume parsers extract candidates' skills from their resumes, and newer technologies can generate free-form text about candidates.[3] These applications raise new concerns, because there is growing evidence that AI-generated text can and often does reflect societal biases [45, 34]. Discussion of bias in these types of systems are beyond the scope of this essay. The important point here is that SPTs are ill-suited for detecting discrimination in these types of AI systems. SPTs do not naturally generalize beyond binary selection decisions, and identifying discrimination in these applications will require a more nuanced approach.

2.2 SPTs as fairness metrics: a prospective view

When considering retrospective liability, a "four-fifths rule" is a poor test for determining whether a practice caused a discriminatory effect because it is simply too crude a measure. Even more robust SPTs that take into account statistical significance have significant limitations. Thus, although SPTs can be helpful tools, when they are applied rigidly, they lose sight of important context and nuance that is relevant to determining whether a practice warrants further scrutiny.

When computer scientists or practitioners invoke a "four-fifths rule" or other SPTs as fairness metrics, they are relying on those statistical measures to guide model development prospectively. Used in this way, additional concerns come into play. The crucial inquiry is no longer "are these tools appropriate for diagnosing discriminatory systems?" but "do they create the right incentives for developing fair models?" The concern is that developers will focus too narrowly on SPTs, making choices keyed to these metrics, rather than trying to understand why disparities are arising and where substantive unfairness may be affecting the selection process. In other words, they may build models to pass statistical tests rather than looking for models that will actually reduce inequities when implemented in the real world.

Of course, industrial-organizational psychologists have long considered the four-fifths ratio and other statistical tests prospectively, using them to evaluate selection instruments [23]. In these traditional practices, however, the test designer evaluates an instrument with a suite of tests including SPTs to see if it is suitable for deployment. The designer typically seeks a qualitative understanding of the performance of the instrument, and makes judgments whether to adjust it, or to adopt some alternative, by weighing validity, adverse impact, and other job-related considerations.

Novel algorithmic techniques to automatically enforce a "4/5 rule" short-circuit this process, by substituting qualitative judgments with a mechanically enforced rule. Instead of a person with substantive expertise making a reasoned decision about the trade-offs from one instrument to another, the developer pre-specifies trade-offs to optimize for compliance with a four-fifths ratio. This automated optimization process reduces search costs, but comes at the cost of qualitative understanding. The developer may have little intuition as to what alternative models the algorithm failed to produce [5]. If the developer can precisely specify their objective (for example, the potential trade-off they are willing to make between predictive accuracy and differences in selection rates), then this lack of intuition may have little practical impact. But to the extent that a developer is unable to completely specify their preferences (for example, that the resultant model refrains from heavy reliance on a candidate's place of education), the developer has little control over the resultant model. Below, we highlight several substantive technical limitations of relying on SPTs prospectively.

Limited measures of performance and bias. Whether or not a model satisfies the four-fifths "rule" has little bearing on whether it accurately predicts outcomes. A model that outputs purely random

[3] See, for example, Affinda (https://www.affinda.com/resume-parser) and parsio (https://parsio.io/).

predictions for two demographic groups has no predictive validity, yet it satisfies a 4/5 rule. AI models are typically built to predict "labels" (often denoted by Y), which are simply values for the target outcome of interest that the model is trying to predict. Examples of labels used in employment models include employee retention, job performance measures like sales numbers, and psychometric traits [41]. A developer seeks to predict the correct label for each individual—for example, will the person still be employed after two years—using available data about that person—i.e. their features.

In order to build a model, a developer uses a dataset—the training data—that contains information about numerous individuals. For each candidate, the data contains information about features (X) as well as information about the outcome of interest—i.e. the label Y (employed after two years / no longer employed after two years) for that individual. In employment settings, this dataset often includes each candidate's demographic information (A), like race and sex, as well. In this notation, the developer's goal is to build a model that, given a new candidate's features X, generates a prediction ($\hat{Y}$) of the label for that candidate.

Recall that the four-fifths ratio, and SPTs more generally, test for disparities in selection rates. Selection rates depend only on predictions ($\hat{Y}$) and demographic characteristics (A), since they only measure the rates at which members from different demographic groups receive positive predictions. Crucially, they do not depend on labels (Y)—i.e. the actual value of the target of interest. This limitation is inherent to ex post evaluation: a model developer simply cannot observe outcomes for candidates who were not selected. Thus, while labels (Y) are key to model *development*, they cannot be observed across all candidates after model *deployment*.

Measures of predictive validity, or the accuracy of a model, typically involve comparisons between Y and $\hat{Y}$. The closer $\hat{Y}$ is to Y, the greater the predictive validity. Additionally, many widely used notions of test bias from the psychology literature depend on Y, $\hat{Y}$, and a demographic attribute A [39]. We focus on two in particular: subgroup calibration, which measures whether a given prediction corresponds to similar outcomes of interest for members of different demographic groups, and differential validity, which measures discrepancies in predictive accuracy across groups. These are both important notions that describe whether an assessment unfairly favors one group over another, and are typically measured using these three attributes: Y, $\hat{Y}$, and A [4, 8, 30]. These measures provide a more nuanced understanding of how a model performs for different demographic groups, and have been commonly used throughout both the psychology and computer science literatures [4, 30]. SPTs cannot capture these important concepts simply because they lack information about Y. We build on this observation in Section 3.

Data representativeness. Algorithm developers sometimes use prospective testing to claim that their models pass SPTs, often the four-fifths "rule" [7, 41]. Without further context, however, this claim is ill-defined: whether or not a model passes an SPT depends crucially on the data on which it is evaluated. A model may pass an SPT on one source of data and fail it on another. Thus, we cannot conclude that a model in isolation either passes or fails an SPT; instead, we can only evaluate whether a model passes *with respect to a particular dataset*. And as a result, evaluating models requires examining not only the results of SPTs, but also the dataset to which they were applied.

In the litigation context, in order to determine whether a prima facie case exists, courts scrutinize hiring decisions in retrospect. The relevant dataset consists of actual applicants to the position, and the decisions made about them. We can determine whether hiring practices satisfied an SPT by examining the outcomes they produced for real people. Algorithm developers, however, are often interested in prospective, as opposed to retrospective, analyses. They want to determine whether a model *will* pass an SPT when it is deployed, not whether it has already done so. In order to make this assessment, an algorithm developer effectively needs to guess what the distribution of future candidates will be, evaluate the model based on this guess, and hope that the guess wasn't too far off when deploying the model. In practice, a developer might use a dataset comprised of past applicants or collect data from a population that they believe to be representative—i.e. their best guess of what the actual applicant pool will look like.

Importantly, the dataset must be representative of the true population in all respects. Finding a dataset that is (for example) demographically representative does not guarantee that the data are representative for other attributes (e.g., education level or work history) that are relevant to the model's predictions. If well-qualified members from some demographic group are over-represented in the dataset, a model may pass an SPT on that dataset but fail to achieve it in practice when the

prevalence of qualified applicants drops. Similar challenges exist for techniques like propensity score reweighting [42] designed to make a dataset representative: while they can re-weight or modify a dataset to be representative along a few known axes, they cannot in general enforce representativeness on all attributes. Moreover, data distributions vary with geography and time: passing an SPT with data from a particular location and time provides no guarantees in other contexts. This dramatically complicates model evaluation for firms seeking to create off-the-shelf models.

Because firms have considerable discretion in selecting the dataset on which to evaluate a model, it is difficult to know if they have done so in good faith. Regulations that rely on prospective auditing can create incentives to curate datasets that make it "easier" for a model to pass an SPT. If a firm is worried that their model under-selects applicants from a particular demographic group, they may simply add more qualified applicants from that demographic group to their dataset, thereby increasing the group's measured selection rate on that dataset. For SPTs that measure statistical significance in addition to effect size, firms may rely on smaller datasets since these are less likely to lead to statistically significant results. Of course, simply changing the dataset on which a model is evaluated will not affect its potential for adverse impact in practice; it simply makes the model *appear* (for the sake of prospective analysis) less discriminatory.

One tempting response to the problems introduced by data collection is to attempt to centralize collection. If a third party (e.g., a regulator) collects and maintains data, firms will lose their ability to manipulate datasets used for SPTs. This approach faces a major hurdle: datasets used to evaluate a predictive model must contain exactly the information required as input to that model. A model that makes predictions based on recorded video interviews requires a dataset containing such interviews. A model that makes predictions based on questionnaires requires a dataset of responses to questionnaires. Thus, the dataset used for a firm's model must be specific to the firm in question; a regulator cannot simply collect a common dataset to be used by all firms.

Determining the relevant pool. The problem of data representativeness is compounded in situations in which the applicant pool is affected by employer behavior, or is otherwise difficult to define. For example, the Guidelines recognized that if an employer has discouraged minority or female candidates from applying, differences in passage rates on a screening test may not accurately measure the overall effect of the employer's selection practices. And conversely, employers who engage in special recruiting efforts to increase the number of minority or female applicants should not necessarily have their practices judged solely by disparities in selection rates. Similarly, when assessing some algorithmic hiring tools, it may be difficult to determine who should be considered part of the candidate pool. Consider an algorithm designed to search a platform's inventory of candidate profiles and recommend to a recruiter the 10 best matches for their position. How should we think about the relevant candidate pool in this case? The pool cannot be defined by who submitted an application, because no one did in this type of situation. So should the relevant pool be the set of all candidates on the platform? Just those who work in the same industry? Or only consider those with appropriate qualifications? Whether or not a model passes an SPT will depend heavily on how we construct this baseline. When used for prospective analysis, an SPT gives a fair amount of flexibility, offering firms an opportunity to choose a baseline that increases their likelihood of passing the test.

3 SPTs and Algorithmic Governance

A four-fifths ratio, or SPTs more generally, have significant limitations when used prospectively as an optimization metric. Relying solely on such measures ignores other relevant metrics that may be important for fairness, such as differential validity. At the same time, it risks creating incentives for gaming, encouraging developers to make choices designed to satisfy the tests rather than seeking substantive understanding of the sources of unfairness and addressing them directly.

Given these limitations of a four-fifths rule and SPTs more generally, what role should they play in legal and policy efforts to prevent algorithmic discrimination? We argue here that SPTs remain useful in the litigation context when examining practices retrospectively, so long as they are not applied rigidly or mechanically. However, when it comes to prospective testing algorithms as part of an auditing requirement, we propose alternatives that will more effectively incentivize fair models.

3.1 Litigating discrimination retrospectively

As explained above, the "four-fifths rule" emerged as a rough indication of when a plaintiff had established a prima facie case of disparate impact, warranting further legal scrutiny of an employer's practice. As explored above, a "four-fifths rule" has significant limitations and courts for the most part have recognized that, if mechanically applied, it is far too crude a measure of possible discrimination. More robust SPTs that test for statistical significance also have limitations, but in the litigation context, they provide a reasonable place to start the analysis. Because the inquiry is inherently retrospective and outcomes cannot be observed for the entire population, examining differences in selection rates offer a first cut at the problem. So long as they are interpreted with nuance and attention to context, SPTs can usefully draw attention to situations that warrant greater legal scrutiny [10].

Despite their limitations, SPTs have some desirable properties. For one, SPTs do not require knowledge of actual outcomes across the population—information that is simply unavailable in some circumstances. Unlike measures such as differential item functioning or error rate disparities, SPTs do not measure validity. While that is a limitation in some respects, it has the advantage that SPTs will not be distorted by inaccurate or biased labels. For example if a firm that discriminated in the past seeks to predict hiring decisions using data about prior decisions, its earlier discrimination will be reflected in how outcomes are labeled. In other words, the labelled data will reflect the results of a biased process, not an objective measure of who would have been the best hires. And to the extent that these labels are systematically biased against one demographic group, measures that take labels into account will fail to detect discrimination. In contrast, SPTs will be unaffected by biased labels because they make no attempt to take labels into account. As a result, SPTs can serve as a check against poor or biased measures of outcomes.

In this sense, SPTs can be viewed as aspirational in nature. Instead of assessing the world as it is by accepting background inequities that may contribute to disparate outcomes, SPTs steer attention towards a world where there are no significant differences between different demographic subgroups. When such differences appear, it triggers questions—specifically, by requiring employers to show that the differing outcomes are justified because they accurately reflect relevant differences between candidates. Such legal scrutiny creates incentives to examine practices that contribute to inequity and to push for expanded opportunities for those who have historically been underrepresented.

Finally, SPTs can create some benefits by pressuring firms to search for equally accurate models with minimal adverse impact. While the computer science literature has frequently explored trade-offs between reducing adverse impact and validity, recent research indicates that very often multiple models exist that have similar performance, but differ in the degree of disparate effects they produce across groups [5, 36]. Because models with very similar accuracy can vary dramatically in disparity rates, a litigation framework that looks to SPTs can encourage firms to seek alternative models of comparable performance that minimize adverse impact.

3.2 Auditing algorithms prospectively

In recent years, it has become increasingly common for researchers and policy-makers to call for audits and impact assessments as ways to address algorithmic discrimination [2, 16, 58]. An ordinance requiring a pre-deployment audit of any employment selection algorithm was passed in New York City [9], and numerous proposed laws have included similar provisions [e.g. 15, 14]. Auditing requirements, however, are typically vague about what the auditing should entail. In other words, they typically lack detail about what types of analyses should be included in a required audit. In the absence of specific direction, some have turned to a 4/5 rule as a pre-deployment test of whether an algorithm discriminates. Given its limitations, we suggest here how pre-deployment audits might go beyond SPTs, maintaining the protections they provide while addressing some of their deficiencies. The goal of auditing should be to provide more meaningful information and incentivize the creation of fair algorithms.

In Section 2, we explained how some measures of model performance such as subgroup calibration and differential validity cannot be performed on data generated by a model deployed in the real world. These measures require information about labels (Y), which are unavailable when a model is actually deployed to make decisions due to selective labels [33].

But this limitation does not apply to prospective testing during model development. A model developer generally has a dataset that does contain information about actual outcomes (Y) for the

training data in addition to predicted outcomes ($\hat{Y}$) and demographic information (A). For example, a developer building a model to predict retention necessarily has historical retention data (Y), which is the objective that the model is designed to predict. Label information (Y) is thus available to a model developer before deployment, but is generally unobservable for the actual applicant population because not all applicants will be hired. In other words, a model developer can measure properties *prospectively* that are impossible for a regulator to measure after the fact.

There are two key challenges in using prospective evaluation to assess the potential for a model to discriminate. First, labels are rarely objective. For example, suppose a model developer seeks to predict Y = performance reviews. To the extent that performance reviews in the training set systematically and unfairly undervalue members of a demographic group, a model can reproduce these patterns. This problem, often referred to as "label bias," is in general difficult to identify in datasets. Below, we discuss heuristics developers can use to look for it, with the caveat that technical methods alone cannot fully capture label bias.

Second, prospective evaluation suffers from a data-dependence problem: a firm cannot be sure that the datasets on which they build and evaluate models will be representative of the true data distribution. If the candidates who apply to a position differ dramatically from those on whom the model was developed and evaluated, a firm cannot guarantee that its conclusions drawn from prospective testing will hold in practice.

This analysis highlights the trade-offs between ex ante and ex post evaluation. Ex ante, we can test for a broader range of relevant properties, particularly those that require information about labels. But the conclusions we draw are only valid insofar as the labels are unbiased and the true candidate distribution resembles the dataset used for ex ante evaluation. In contrast, we perform ex post evaluation on the actual set of applicants, meaning we do not need to guess what the candidate distribution will be. However, the lack of labels makes it impossible to evaluate important measures of test bias, such as differential validity.

3.3 A concrete technical proposal

Given these trade-offs, SPTs likely should remain a part of the auditing process; however, we propose three additional concrete measures that should be part of pre-deployment audits in order to take advantage of the information available during model development and to fill the gaps if only SPTs are considered.

1. **Predictive validity.** For models with binary outputs, this would include measures of error rates like precision and recall. For models with continuous outputs, firms should report global performance measures such as ROC-AUC (which aggregates predictive quality across all possible thresholds) in addition to performance measures like precision and recall that are specific to a threshold. Crucially, the thresholds used to report model performance should reflect the intended thresholds for use in practice. If a model is to be used for ranking instead of classification, firms can use performance measures from the information retrieval literature designed to measure ranking systems [43].

2. **Differential validity.** Firms should report validity disaggregated by demographic groups. Note that this is substantially different from what SPTs measure: a model may select members of different demographic groups at the same rate but have far worse predictive validity on one group than another, leading to negative downstream consequences. While some psychologists have noted that testing for differential validity has historically been uncommon in employment settings [37], recent work in machine learning has demonstrated that when models are trained on datasets with disparate amounts of data from different demographic groups, they can exhibit large performance disparities in practice [8, 31].

3. **Subgroup calibration.** For models with continuous outputs, firms should be required to report whether, conditioned on receiving the same predictions, members of different demographic groups exhibit differences in their labels. In the IO psychology literature, this is often known as "test bias," and researchers have developed a variety of measures to quantify it [4, 26, 39]. While many of these methods are based on regression, as opposed to the more sophisticated machine learning methods deployed today, the computer science and

statistics literatures propose alternative frameworks to assess miscalibration, and other firms have developed tools to report it.[4]

Beyond these quantitative measures, firms can and should take additional steps to guard against label bias. Firms should document their label definitions, data sources, and model development process. The computer science literature contains a variety of documentation tools that have been deployed across a wide range of contexts [24, 38]. Moreover, firms can attempt to identify label bias by measuring whether a dataset admits *differential prediction*, which occurs when the most predictive model for one demographic group differs dramatically from the most predictive model for another [37]. Differential prediction provides evidence that similar individuals (in terms of their features X) receive different labels Y, which can indicate label bias. However, differential prediction is typically measured in the context of linear models; adapting these measures to more general classes of machine learning models remains an area of active research [5, 32, 37].

While these additional measures go a long way towards addressing some of the limitations of SPTs, they do not solve all of the challenges inherent in relying on pre-deployment audits to identify discriminatory models. For this reason, "passing" an audit should not shield a firm from later inquiries about whether its algorithm discriminates.[5] If an audit conferred legal immunity, firms would be incentivized to manipulate the audit process, undermining their utility and potentially allowing discriminatory algorithms to escape scrutiny.

One significant limitation is inherent to prospective evaluation: in order to evaluate a model before it is deployed a firm must effectively guess the applicant distribution in order to collect a representative dataset. To the extent that they guess wrong, conclusions derived from an audit may be invalid. Even with the best intentions, a firm may simply not know what the applicant pool will look like. And if audit requirements guaranteed legal immunity, firms would have an incentive to curate a dataset that yields the desired results rather than to select the most likely representative dataset to accurately diagnose the risks of bias. Data-dependence is thus a key limitation on how useful audits can be. Designing statistical techniques to determine whether the dataset used for an audit is sufficiently representative of the actual applicant pool in hindsight is an important direction for future work.

In order for audits to be effective in preventing biased algorithms, they must be conducted with due diligence and in good faith. Deploying firms and the contexts in which algorithms operate are heterogeneous enough that audits cannot be standardized [2]. Even where precise technical specifications are possible, firms retain a great deal of latitude to make choices, including the relevant candidate pool, which outcomes to report, thresholds to set, and the exact metrics they choose to report. This discretion creates challenges for designing a meaningful audit process. Audits could be conducted by the firm itself, a third-party, or a government regulator, and each of these approaches has advantages and drawbacks. However these issues are resolved as a matter of regulatory design, at a minimum, firms should be required to document and disclose the choices made in conducting the audit in order to enhance its reliability and trustworthiness.

For all these reasons, the requirements for pre-deployment audits should not be seen as metrics that guarantee that models will not discriminate. Instead, they should be crafted with an eye to creating incentives for firms to understand the risks of bias and to make choices that minimize those risks, while increasing the transparency of the model building process.

4 Conclusion

Firms have relied on compliance with a "four-fifths rule" to develop and optimize models for algorithmic hiring in the hopes of avoiding legal liability. The four-fifths ratio was already a poor test for determining whether practices warrant close legal scrutiny. Its use prospectively as a measure of fairness to shape how algorithms are built raises additional concerns. And yet, as firms and vendors increasingly reference the so-called rule, it risks becoming the de facto standard for legal compliance, creating incentives to comply with what appears to be the letter of the law without

[4] For example, Meta's Fairness Flow considers subgroup miscalibration: `https://ai.facebook.com/blog/how-were-using-fairness-flow-to-help-build-ai-that-works-better-for-everyone/`

[5] The only currently enacted legislation in the US requiring audits for hiring algorithms does not foreclose a later legal challenge if the algorithm turns out to be discriminatory in operation [9].

addressing substantive questions of discrimination. In this work, we have detailed the various ways in which the four-fifths ratio is inadequate as a retrospective test for liability.

Similarly, prospective evaluation of algorithmic employment assessments can and should go beyond the four-fifths ratio and closely-related statistical parity tests. Firms can leverage the additional information available during model development to provide a more nuanced picture, which we have detailed above, of how an assessment performs for members of different demographic groups. Similarly, to the extent that regulators impose auditing requirements on firms, these audits should include a more comprehensive set of tests. While discrimination cannot be reduced to a suite of statistical tests, developers and regulators have at their disposal multiple tools to assess the performance and fairness of algorithmic assessments beyond the simple four-fifths ratio.

References

[1] Federal register. 43(166), 1978.

[2] Solon Barocas, Anhong Guo, Ece Kamar, Jacquelyn Krones, Meredith Ringel Morris, Jennifer Wortman Vaughan, W Duncan Wadsworth, and Hanna Wallach. Designing disaggregated evaluations of ai systems: Choices, considerations, and tradeoffs. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 368–378, 2021.

[3] Jason R Bent. Is algorithmic affirmative action legal. *Geo. LJ*, 108:803, 2019.

[4] Christopher M Berry. Differential validity and differential prediction of cognitive ability tests: Understanding test bias in the employment context. *Annu. Rev. Organ. Psychol. Organ. Behav.*, 2(1):435–463, 2015.

[5] Emily Black, Manish Raghavan, and Solon Barocas. Model multiplicity: Opportunities, concerns, and solutions. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 850–863, 2022.

[6] Anthony E Boardman and Aidan R Vining. The role of probative statistics in employment discrimination cases. *Law & Contemp. Probs.*, 46:189, 1983.

[7] Miranda Bogen and Aaron Rieke. Help wanted: An examination of hiring algorithms, equity, and bias. Technical report, Upturn, 2018.

[8] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

[9] New York City. Nyc administrative code ch. 5 § 20-870.

[10] Equal Employment Opportunity Commission. Select issues: Assessing adverse impact in software, algorithms, and artificial intelligence used in employment selection procedures under title vii of the civil rights act of 1964. 2023. URL `https://www.eeoc.gov/laws/guidance/select-issues-assessing-adverse-impact-software-algorithms-and-artificial`.

[11] Equal Employment Opportunity Commission, Civil Service Commission, et al. Uniform guidelines on employee selection procedures. *29 C.F.R. §1607*, 1978.

[12] United States Congress. 42 U.S. Code § 2000e-2. URL `https://www.law.cornell.edu/uscode/text/42/2000e-2`.

[13] United States Congress. Civil rights act of 1964, 1964.

[14] United States Congress. American data privacy and protection act, h.r. 8152, 2022.

[15] United States Congress. Algorithmic accountability act h.r. 6580, 2022.

[16] Sasha Costanza-Chock, Inioluwa Deborah Raji, and Joy Buolamwini. Who audits the auditors? recommendations from a field scan of the algorithmic auditing ecosystem. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1571–1583, 2022.

[17] United States Supreme Court. Albemarle paper co. v. moody, 422 u.s. 405 (1975), 1975.

[18] United States Supreme Court. Castaneda v. partida, 430 u.s. 482, 496, 1976.

[19] United States Supreme Court. New york city transit authority v. beazer, 440 u.s. 568, 584, 1979.

[20] United States Supreme Court. Watson v. fort worth bank & trust, 487 u.s. 977, 1988.

[21] United States Supreme Court. Ricci v. destefano, 557 u.s. 557, 587, 2009.

[22] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 259–268, 2015.

[23] Society for Industrial, Organizational Psychology (US), and American Psychological Association. Division of Industrial-Organizational Psychology. *Principles for the validation and use of personnel selection procedures*. American Psychological Association, 2018.

[24] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

[25] Zach Harned and Hanna Wallach. Stretching human laws to apply to machines: The dangers of a" colorblind" computer. *Fla. St. UL Rev.*, 47:617, 2019.

[26] Ben Hutchinson and Margaret Mitchell. 50 years of test (un) fairness: Lessons for machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 49–58, 2019.

[27] Rick Jacobs, Kevin Murphy, and Jay Silva. Unintended consequences of eeo enforcement policies: Being big is worse than being bad. *Journal of Business and Psychology*, 28:467–471, 2013.

[28] Faisal Kamiran and Toon Calders. Classifying without discriminating. In *2009 2nd international conference on computer, control and communication*, pages 1–6. IEEE, 2009.

[29] Pauline T Kim. Race-aware algorithms: Fairness, nondiscrimination and affirmative action. *Cal. L. Rev.*, 110:1539, 2022.

[30] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *Innovations in Theoretical Computer Science*, 2017.

[31] Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R Rickford, Dan Jurafsky, and Sharad Goel. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689, 2020.

[32] Finn Kuusisto, Vitor Santos Costa, Houssam Nassif, Elizabeth Burnside, David Page, and Jude Shavlik. Support vector machines for differential prediction. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2014, Nancy, France, September 15-19, 2014. Proceedings, Part II 14*, pages 50–65. Springer, 2014.

[33] Himabindu Lakkaraju, Jon Kleinberg, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. The selective labels problem: Evaluating algorithmic predictions in the presence of unobservables. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 275–284, 2017.

[34] Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. Towards understanding and mitigating social biases in language models. In *International Conference on Machine Learning*, pages 6565–6576. PMLR, 2021.

[35] Zachary Lipton, Julian McAuley, and Alexandra Chouldechova. Does mitigating ml's impact disparity require treatment disparity? *Advances in neural information processing systems*, 31, 2018.

[36] Charles Marx, Flavio Calmon, and Berk Ustun. Predictive multiplicity in classification. In *International Conference on Machine Learning*, pages 6765–6774. PMLR, 2020.

[37] Michael A Mcdaniel, Sven Kepes, and George C Banks. The uniform guidelines are a detriment to the field of personnel selection. *Industrial and Organizational Psychology*, 4(4):494–514, 2011.

[38] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229, 2019.

[39] James L Outtz. *Adverse impact: Implications for organizational staffing and high stakes selection*. Taylor & Francis, 2010.

[40] Manish Raghavan and Solon Barocas. Challenges for mitigating bias in algorithmic hiring. *Brookings Institution*, 2019.

[41] Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 469–481, 2020.

[42] Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

[43] Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. *Introduction to information retrieval*, volume 39. Cambridge University Press Cambridge, 2008.

[44] Elaine W Shoben. Differential pass-fail rates in employment testing: Statistical proof under title vii. *Harvard Law Review*, pages 793–813, 1978.

[45] Eric Michael Smith, Melissa Hall, Melanie Kambadur, Eleonora Presani, and Adina Williams. "i'm sorry to hear that": Finding new biases in language models with a holistic descriptor dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9180–9211, 2022.

[46] Marion G Sobol and Charles J Ellard. Evaluating the four-fifths rule vs. a statistical criterion for the determination of discrimination in employment practices. *Lab. Stud. J.*, 10:153, 1985.

[47] Kevin Tobia. Disparate statistics. *The Yale Law Journal*, pages 2382–2420, 2017.

[48] District of Columbia Circuit United States Court of Appeals. Frazier v. consol. rail corp., 851 f.2d 1447, 1451, 1988.

[49] Eighth Circuit United States Court of Appeals. Mems v. city of st. paul, 224 f.3d 735, 740, 2000.

[50] First Circuit United States Court of Appeals. Fudge v. city of providence fire dep't,766 f.2d 650, 658 n.10, 1985.

[51] First Circuit United States Court of Appeals. Jones v. city of boston, 752 f. 3d 38, 2014.

[52] Ninth Circuit United States Court of Appeals. Clady v. county of los angeles, 770 f.2d 1421, 1428, 1985.

[53] Second Circuit United States Court of Appeals. M.o.c.h.a. soc'y, inc. v. city of buffalo, 689 f.3d 263, 274, 2012.

[54] Seventh Circuit United States Court of Appeals. Bew v. city of chicago, 252 f. 3d 891, 2001.

[55] Sixth Circuit United States Court of Appeals. Isabel v. city of memphis, 404 f. 3d 404, 2004.

[56] District of Columbia United States District Court. Stagi v. nat'l r.r. passenger corp., 391 fed. appx. 133, 138, 2010.

[57] District of New Jersey United States District Court. Vulcan pioneers, inc. v. new jersey dep't of civ. serv., 625 f. supp. 527, 544, 1985.

[58] Briana Vecchione, Karen Levy, and Solon Barocas. Algorithmic auditing and social justice: Lessons from the history of audit studies. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–9, 2021.

[59] Elizabeth Anne Watkins, Michael McKenna, and Jiahao Chen. The four-fifths rule is not disparate impact: a woeful tale of epistemic trespassing in algorithmic fairness. *arXiv preprint arXiv:2202.09519*, 2022.

[60] Christo Wilson, Avijit Ghosh, Shan Jiang, Alan Mislove, Lewis Baker, Janelle Szary, Kelly Trindel, and Frida Polli. Building and auditing fair algorithms: A case study in candidate screening. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 666–677, 2021.

[61] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*, pages 1171–1180, 2017.

A Legal background

A.1 Origins of the "four-fifths rule"

When the Civil Rights Act of 1964 outlawed discrimination in employment, uncertainty arose about the legality of pre-employment tests which were widely used by employers. These tests were not necessarily intentionally discriminatory, which would have clearly violated the law as a form of disparate treatment. Nevertheless, they often had a disparate racial impact on hiring. Different federal agencies, each having some enforcement responsibilities, developed different guidelines regarding pre-employment tests. In an effort to produce a consistent government position, the principal agencies involved in enforcing employment discrimination laws (Equal Employment Opportunity Commission, Civil Service Commission, Department of Justice, Department of Labor) issued the 1978 Uniform Guidelines on Employee Selection Procedures (Guidelines) [11].

The Guidelines confirmed the basic framework to be followed in evaluating employee selection procedures. Where a practice had an adverse impact on protected groups, the agencies would consider it discriminatory unless justified [11]. The Guidelines then explained in considerable technical detail how a test could be validated under existing professional standards established by industrial psychologists.

Importantly, if a test or procedure had no adverse impact, the agencies would not require validity studies [11]. Thus, the question of what constituted an adverse impact became salient. The Guidelines explained that if the ratio of selection rates between two groups was less than four-fifths, it would "generally be regarded by the Federal enforcement agencies as evidence of adverse impact, while a greater than four-fifths rate will generally not be regarded by Federal enforcement agencies as evidence of adverse impact" [11].

The four-fifths ratio was never intended to be a rule of law, but rather a "rule of thumb." It offered a "practical device" to guide the enforcement priorities of the relevant agencies, focusing their attention on practices that caused "serious discrepancies" in hiring and promotion rates [1]. The Guidelines specifically disclaimed application to the resolution of individual complaints alleging discrimination [11].

Even for the enforcement agencies, the "rule" was not controlling. The Guidelines recognized that, rigidly applied, the four-fifths ratio was both under- and over-inclusive. Smaller differences in selection rates could suffice to show adverse impact when requirements of statistical and practical significance were met, and larger differences might not constitute adverse impact when the sample size was small [11]. The Guidelines further recognized that the context mattered. Special recruitment efforts might increase applicants from disadvantaged groups; discriminatory action might discourage them [11]. In either case the pool of applicants would change in ways that would affect the selection ratio.

Because a finding of adverse impact triggered the requirement of validation and the risk of government scrutiny, the four-fifths ratio became a focal point of attention. It was immediately the subject of criticism by scholars and advocates on all sides, who argued that it was highly problematic if deployed as a rule for identifying discrimination [see, e.g., 6, 44, 46]. Because validating a test under the Guidelines was technically complex and costly, employers had strong incentives to try to avoid triggering scrutiny in the first place. After the 1970s, however, the federal government's efforts to combat systemic employment discrimination receded, and the role of the four-fifths ratio in guiding agency discretion became less salient.[6]

A.2 Proving disparate impact in practice

In the courts, most of the law around disparate impact liability evolved through cases brought by civil rights groups or private litigants. Cases alleging disparate impact discrimination entail a three-step analysis [12, 17]. First, plaintiffs must establish a prima facie case, usually by producing statistical evidence that shows an employer practice disproportionately screens out a protected group. Second, the employer has the opportunity to defend its practice by showing it is job related and consistent with business necessity [13, 12], or "validating" it in the terminology of the Guidelines. Even if it succeeds in doing so, plaintiffs may nevertheless prevail by pointing to a less discriminatory alternative that would meet the employer's business needs [12]. Thus, as with agency enforcement decisions, an important first step is deciding whether there is sufficient evidence—a prima facie case—to warrant further legal scrutiny.

Although some have suggested that a four-fifths rule should apply, courts have generally not adopted it as the test for establishing a prima facie case of disparate impact. The Supreme Court specifically noted that "the rule has been criticized on technical grounds . . . and it has not provided more than a rule of thumb for the courts" [20]. Federal courts of appeals have similarly refused to treat selection ratios below four-fifths as the legal test of disparate impact [51, 56, 52]. While a selection ratio that falls below that threshold *can* be sufficient to establish a prima facie case [57, 53], it does not always do so, particularly when sample sizes are small [49, 48, 50]. On the other hand, selection ratios above that cutoff do not automatically absolve an employer [54, 55, 51]. Simply put, a selection ratio below 4/5 is neither necessary nor sufficient for a finding of disparate impact under current law.

In the litigation context, the role of the prima facie case is to determine whether there is sufficient evidence of a disparate impact to warrant requiring the employer to defend its employment practice. This inquiry is a retrospective one. It asks whether a particular employment practice, though facially neutral, systematically disadvantaged a marginalized group, and therefore requires justification. Because there is some inevitable randomness in any process, a key question is whether the employer's practice caused the observed difference in selection rates between groups, or whether those differences could have occurred by chance [21, 19, 51].

One problem with relying on a "four-fifths rule" to identify discrimination is that simply looking at the selection ratio does not consider the statistical significance of the effect it is trying to measure. Concretely, consider two firms that each have selection rates of 30% and 20% for men and women respectively. In both cases, the ratio between the selection rates is 0.2/0.3 or 0.67, which is less than 0.8, or four-fifths. In other words, both firms would be considered in violation of a four-fifths rule. Suppose, however, that Firm 1's applicant pool contained 100 men and 100 women (of which it hired 30 men and 20 women), while Firm 2's applicant pool contained 10 men and 10 women (of which it hired 3 men and 2 women). The disparity in selection rates is clearly more meaningful for Firm 1 than Firm 2, even though they have the same selection ratio. If Firm 2 happened to hire one more woman and one fewer man, it would have an adverse impact against men instead of women according to the four-fifths rule. In technical terms, a four-fifths rule considers only effect size (how far apart the selection rates are) and not statistical significance (the likelihood of observing the same results by random chance).

Recognizing this limitation, courts have looked to a variety of tests to determine whether an observed difference in selection rates is statistically significant [18]. For example, they may use formal statistical tests such as Fisher's exact test and the chi-squared test to ask whether an observed difference in selection rates is statistically significant using a conventional cutoff like 10, 5 or 1% [51].

[6] In a recently issued technical assistance document, the EEOC reiterated that the four-fifths is merely a rule of thumb that is not always appropriate to rely on and should not substitute for a test of statistical significance [10].

These tests seek to determine whether observed disparities in the selection rates of different groups could have arisen by chance, or were more likely caused by the challenged employment practice.

In addition to statistical significance, some courts also ask whether the magnitude of differences in selection rates is meaningful—i.e. whether it is practically significant. To measure practical significance, courts may refer to a four-fifths ratio, but often use other measures [47].

As discussed below, there are other ways to compare the impact of a selection procedure on different groups, but the law has traditionally focused on SPTs, although not, as commonly assumed, using a four-fifths ratio as a cutoff. To be clear, the outcome of a statistical test does not establish whether discrimination occurred. Rather, it is the first step in the legal process of determining whether a given practice is considered discriminatory. Statistical evidence is used to establish a prima facie case, which then shifts the burden to the employer to justify the practice. If the employer can demonstrate the validity and necessity of its practice, it generally will avoid liability for discrimination.[7] However, facing a prima facie case is costly to an employer, who must mount a legal defense and gather evidence supporting its practices. As such, employers face strong incentives to avoid legal jeopardy in the first place by tailoring their practices to avoid a prima facie case [41, 40]. The legal standard for establishing a prima facie case will thus shape how algorithmic hiring tools are developed.

B The "four-fifths rule" as a fairness metric

As machine learning techniques are applied to decision-making in social domains like employment, concerns have grown that predictive algorithms may be discriminatory and unfair. Computer scientists and practitioners have sought methods to ensure that models are fair, and some have looked to the law for guidance [22, 61].

As a result, a four-fifths "rule" has sometimes been adopted as a common metric by which the fairness of models are evaluated. Much of the academic research that makes reference to the four-fifths "rule" is not specific to the employment context; instead, compliance with a four-fifths ratio is seen as one possible property for characterizing model fairness across varying applications and contexts.

When developers invoke a statistical test like a four-fifths rule, its function is different than in the legal context where the focus is retrospective. Instead of measuring the amount of impact that has occurred after the deployment of a test, it is used as a way of defining fairness/nondiscrimination prospectively when building models. Rather than a way of examining the connection between a practice and an observed disparity, it is used as an optimization objective.

A common goal of this research is to develop machine learning algorithms to automatically train models that comply with a fairness metric, such as a four-fifths rule. To a first approximation, we can think of traditional machine learning algorithms as following the instruction, "find me the model that makes the most accurate predictions on this data." Algorithmically enforcing a four-fifths rule amounts to modifying that instruction to "find me the model that makes the most accurate predictions on this data, subject to the constraint that the ratio of selection rates does not fall below four-fifths."

In the literature, a variety of techniques have been developed to achieve this end. However, many of them directly rely on individual demographic characteristics as an input feature. While there is debate in the legal literature on this point [3, 29], developers are sufficiently concerned that this practice would itself be considered a form of discrimination to limit its practical application in the employment context.

A separate class of techniques, sometimes called "Disparate Learning Processes" or DLPs, also seek to ensure compliance with fairness metrics, but without relying on an applicant's demographic characteristics when making predictions [28, 35]. While some scholars have argued that DLPs could bypass legal concerns about relying on protected characteristics [25], these strategies have yet to gain traction in practice.

A "four-fifths rule" is often invoked by practitioners as well. For example, vendors who build predictive algorithms for use in social domains like consumer finance, housing, employment, and criminal law enforcement sometimes refer to a "four-fifths rule" as a measure of legal compliance.

[7] Even if the employer meets its burden of justification, it might nevertheless be found liable if the plaintiff identifies a less discriminatory alternative that the employer refuses to adopt [17]. Very few cases, however, succeed by following this route.

Others have promoted "toolkits" that offer generalized approaches to ensure fair algorithms across use cases that refer to the four-fifths ratio [see 59]. Many of these also refer to the four-fifths rule as a relevant metric.

As concerns about algorithmic discrimination have moved into the policy sphere, auditing is emerging as an important governance tool [9]. In the absence of well-established auditing standards, once again, some have looked to the four-fifths "rule" as a measure of compliance [60]. Other proposals have suggested the use of pre-certification requirements or licensing standards, but implementation of those regimes will also require reference to a substantive measure of nondiscrimination.

The role of the "four-fifths rule" differs somewhat in practice compared with the automated techniques found in the computer science literature. After developing a machine learning model, firms often run a suite of SPTs that measure differences in selection rates across demographic groups. If significant differences are found, the firm will remove data attributes that contribute to these differences, re-build the model, and repeat until the model passes the tests [7, 41].

While firms claim that this procedure is a good-faith attempt to ensure non-discrimination, it may also be motivated by litigation avoidance: if a firm produces a model that passes the statistical test it believes courts will use, then it may be difficult or impossible for a plaintiff to demonstrate a prima facie case, and the firm can avoid scrutiny. As discussed previously, however, a procedure that satisfies the four-fifths rule of thumb may still warrant further legal scrutiny. The prospective use of the four-fifths rule can thus become a strategy for model developers to minimize legal risk without addressing the substantive harms of potential discrimination. In what follows, we demonstrate the limitations of over-reliance on the four-fifths rule as a test for discrimination in algorithms.