Main takeaway

Consider fairness in 3 dimensions (and different stakeholders): *product*, *policy*, and *implementation*. In *implementation*, consider not only the ML models but also the human labelers.

Fairness On The Ground: Applying Algorithmic Fairness Approaches To Production Systems

Chloé Bakalar, Renata Barreto, Miranda Bogen, Sam Corbett-Davies, Melissa Hall, Isabel Kloumann, Michelle Lam, Joaquin Quiñonero Candela, Manish Raghavan, Joshua Simons, Jonathan Tannen, Edmund Tong, Kate Vredenburgh, Jiejing Zhao
Facebook

ABSTRACT

Many technical approaches have been proposed for ensuring that decisions made by machine learning systems are fair, but few of these proposals have been stress-tested in real-world systems. This paper presents an example of one team's approach to the challenge of applying algorithmic fairness approaches to complex production systems within the context of a large technology company. We discuss how we disentangle normative questions of product and policy design (like, "how should the system trade off between different stakeholders' interests and needs?") from empirical questions of system implementation (like, "is the system achieving the desired tradeoff in practice?"). We also present an approach for answering questions of the latter sort, which allows us to measure how machine learning systems *and* human labelers are making these tradeoffs across different relevant groups. We hope our experience integrating fairness tools and approaches into large-scale and complex production systems will be useful to other practitioners facing similar challenges, and illuminating to academics and researchers looking to better address the needs of practitioners.

1 INTRODUCTION

In recent years many technical approaches have been proposed for ensuring that decisions made by machine learning systems are fair. To date, however, few of these proposals have been stress-tested in real-world systems. A viable approach to achieving fairness in practice requires overcoming a number of challenges that do not similarly constrain theoretical work.

First, as researchers increasingly acknowledge [47, 56], purely technical or statistical approaches to fairness leave unanswered important questions related to ethics and policy. An approach to fairness in practice must have means of clearly surfacing and resolving these tensions, which are not reducible to empirical questions.

Second, it has been identified that some common statistical notions of fairness may lead to unintentional and potentially harmful consequences [11, 26, 38], especially to marginalized groups. This could be even more important to consider in decision systems that may affect millions or even billions of people. Instead, what is needed as a baseline is a statistical approach that makes the costs and benefits of decisions—and who these fall on—explicit. The normative decision about who the costs and benefits *should* fall on then becomes part of discussions of ethics and values, or where relevant, law or policy. By disentangling statistical questions from normative ones, this approach is flexible enough to be applied to a wide range of different systems while ensuring normative decisions are made explicitly, with input from relevant stakeholders and domain experts.

Finally, technical approaches to fairness in machine learning often rely on assumptions that may not be reasonable in practice. For example, popular statistical measures of fairness like equality of odds [23] and calibration [33] assume that measured labels constitute ground truth, which is not always the case. Methods that allow for inaccuracies in measured labels, on the other hand, often assume the true prevalence of labels among groups [17, 28], or the causal process that generated them [31, 36, 39, 44]. These assumptions are challenged when studying human reviewers and machine learning systems in dynamic environments where adversarial actors may try to evade systems intended to constrain them.

This paper presents an example of one team's approach to address these challenges within the context of a large technology company. We are not the first to identify any of these challenges, nor are our proposed solutions—taken individually—entirely novel. However, we believe that our experience integrating fairness tools and approaches into large-scale and complex production systems may be useful to other practitioners facing similar constraints, and illuminating to academics and researchers looking to better address the needs of practitioners.

We begin in Section 2 by surveying the fairness approaches that have seen practical implementation and the challenges preventing other approaches from being widely adopted. Section 3 introduces our holistic approach to assessing fairness in large-scale production systems, including a discussion of how we disentangle related yet distinct notions of fairness to facilitate more actionable analysis and decision-making for non-experts. For example, we explicitly consider and investigate fairness across multiple dimensions: at the *product* level, the *policy* level, and in a product or system's *implementation*. Section 4 details more specifically the primary measurements we use to examine implementation fairness, and why we use, as a baseline, a measurement approach grounded in making the weighing of potential benefits and harms explicit rather than deploying methods where such values decisions are implied and obscured in technical choices.

Section 5 explains how this general approach to implementation fairness is applied to two different types of decisions: those made by machine learning models (Section 5.1) and those made by human labelers (Section 5.2). In the case of machine learning predictions, we determine the system's prioritization of errors (false positives and false negatives) for different groups by measuring the *prevalence at the threshold*. To study the implied prioritization of human labelers we use Signal Detection Theory (SDT) [22], an approach with a long history in psychology, to infer the latent "thresholds" that labelers are applying to labeling tasks affecting different groups. We conclude in Section 6 with a discussion of challenges and open

technical questions that complicate the application of theoretical recommendations in practice.

The following presents a lens, and collection of approaches, by which fairness can be considered in a variety of practical cases, and we emphasize the importance of context and subject matter expertise in determining the approach(es) best suited to address potential fairness-related issues in any particular product or domain.

2 RELATED WORK

2.1 Fairness in practice

Despite the explosion of academic interest in methods for developing fair algorithms, fewer methods have been implemented in production machine learning systems used by governments or private companies (at least, few of these entities have been willing to publicly share their fairness approaches, if they exist).[1]

The fairness-enhancing approaches that have achieved the most practical success seem to be efforts to improve performance by adding training data, especially for underrepresented groups. For example, after Buolamwini and Gebru [10] discovered that IBM's facial gender classification system was performing poorly for dark-skinned people—and dark-skinned women in particular—IBM responded with a system trained on more representative data which reportedly reduced the error rates on dark-skinned women almost tenfold [50].[2] Similarly, when researchers at Jigsaw noticed their comment toxicity classifier was labeling innocuous comments containing identity terms (eg "gay" or "muslim") as toxic, they augmented the training data to contain more neutral phrases with these identity terms, improving the system overall [14].

In the healthcare context, researchers found that training an algorithm to predict the costs a patient would incur as a proxy for healthcare need perpetuated bias against African-American patients, who tended to incur lower costs than their white counterparts, conditioned on the same level of health need [46].[3] In response, the healthcare provider is re-evaluating its prediction practices [37]. In these cases, the solution to unfairness was simply better machine learning; a concern for fairness motivated the investigations, but the resulting changes may have been adopted even if the decision makers were concerned solely with efficiency. The oft-discussed [12, 33] tradeoff between fairness and efficiency did not apply.

Other efforts advance to goals of fairness, accountability, and transparency in practice are more procedural in nature. Proposals like "Datasheets for Datasets" [20] and "Model Cards" [40] seek to make explicit the limitations of ML systems without making prescriptions or recommendations as to how to resolve difficult policy questions. Several companies have also released toolkits to help measure multiple fairness-oriented metrics without making prescriptions about the appropriate metric to use [4, 7, 61]. Recognizing that documentation and metrics on their own will not necessarily lead to meaningful fairness improvements, others have proposed a structured frameworks for internal algorithmic auditing informed by organizational values [53].

Few fairness-minded interventions that aren't purely efficiency-oriented have been publicly discussed, with some notable exceptions. In 2019, researchers at LinkedIn published details of a system used in production to explicitly gender-balance the results returned when a recruiter searches for candidates [21]. Similarly, some vendors of algorithmic hiring assessments attempt to ensure that the data-driven models they sell don't produce outcome disparities with respect to protected characteristics like race and gender, though this may be due in part to legal considerations as opposed to purely ethical ones [8, 51, 55]. Practitioners at Google describe the implementation of a particular fairness metric in production, though they aren't specific about the exact setting in which they are working [6].

2.2 Bias in machine learning predictions

The problem of bias in supervised machine learning models is likely the most studied problem in algorithmic fairness. Numerous fairness criteria have been proposed [5, 41, 60], along with means of learning predictors that satisfy a given criterion.

Despite this plethora of options, few approaches appear to have been implemented in consequential production systems (see 2.1 for a discussion of approaches that have seen practical implementations). We suspect there are two reasons for this. First, as Corbett-Davies and Goel [11] discuss, many fairness criteria, including those that require equal positive classification rates (demographic parity) or equal false positive/negative rates (equality of opportunity [23]), may fail to anticipate all implications to the well-being of the people affected by a model, and can thus inadvertently *harm* marginalized groups. For example, Liu et al. [38] find that, in certain lending situations, equalizing false negative rates by borrowers' race would lead to predictable decreases in the credit scores of certain African American borrowers. Similarly, Hu and Chen [26] apply tools from welfare economics to find that "applying more strict fairness criteria that are codified as parity constraints can worsen welfare outcomes for both groups."

Second, as Kleinberg et al. [32] note: "a preference for fairness should not change the choice of estimator". In other words, it is inappropriate to change a system's *predictions* to achieve any fairness goal, since this will inevitably hurt the usefulness of the predictions for all groups. Instead, a desire for fairness should change how the predictions are used to make decisions. Practitioners, acutely aware of the cost of mistakes in their domain, are less likely to choose fairness solutions that increase these costs unnecessarily.

2.3 Bias in training labels

Compared to bias in predictions, bias affecting the *labels* used in machine learning has received less attention in the algorithmic fairness literature. Many algorithmic fairness metrics (including popular ones like calibration [12, 33] and equality of false positive rates [23]) make reference to the "true" labels in an evaluation set, implicitly assuming these labels faithfully represent the desired

[1] See Holstein et al. [24] for a deeper investigation into the needs of machine learning practitioners that aren't being met by current methods.

[2] Subsequent work on face recognition and classification systems has focused on the ethical problems inherent in these tasks, even when the tasks are performed with accuracy for many groups [52]. This neatly illustrates the distinction between fairness in *implementation* (which concerns the performance of the system for different groups), and fairness in *product* and *policy* design. We return to these different "levels" of fairness analysis in Section 3.

[3] Among the many competing measures of fairness, the researchers chose calibration as the measure "most relevant to the real-world use of the algorithm". We similarly found calibration-based approaches most useful in our applications.

target of prediction.[4] It is common for papers discussing these methods to acknowledge that the labels might be biased (some are even motivated in part by the possibility of label bias), but it's rare for a paper to measure this bias.

Some have proposed methods to remedy possible biases in labels [17, 18, 28]. However, without measurable ground truth to fall back upon, these approaches are left making assumptions about ground truth distributions for different groups that may not be applicable in all cases. A different line of work attempts to identify the causal paths that may lead to label biases, so that the effect of these paths can be nullified [31, 36, 39, 44]. Unfortunately, these approaches are very sensitive to the structure of the causal model used, which in most cases cannot be empirically verified.

Beyond this recent computer science research there is a rich literature in economics [2, 3, 34], statistics [49, 57], and psychology [22, 43] (among other fields) developing methods to measure biases in human behavior. These methods offer hope that we might be able to identify label bias in datasets based on human decisions (for example in hiring, policing, college admissions, natural language and visual classification, etc). In section 5.2, we detail how a method from psychology, Signal Detection Theory, can be adapted to detect potential bias in human-provided labels when a source of ground truth is available.

single centralized unit whose parts retain some internal autonomy

3 A HOLISTIC APPROACH TO FAIRNESS

Because there is no single, agreed-upon definition of fairness, operationalizing fairness at the scale of a large and federated organization requires building a shared language for describing fairness risks and conducting standard analyses, cultivating a broad understanding of values and objectives, and establishing tools, processes, and frameworks to enable teams to make informed decisions about fairness across different contexts [19, 45].

These resources cannot just be technical: imposing statistical definitions of fairness on individual machine learning models by fiat without sensitivity to wider systems and contexts in which they are embedded can backfire, failing to benefit disadvantaged groups and undermining rather than promoting fairness over time. Particularly in dynamic, complex systems like online platforms, individual machine learning models may be the wrong level at which to impose substantive requirements of fairness, so a more holistic approach is required.

We thus consider fairness at three distinct but related levels: *product*, *policy*, and *implementation*. Fairness at the product level relates to normative and descriptive questions about the product, such as: "Are the goals of this product consistent with providing people with fair value and treating them fairly?" and "How should the product trade off between different stakeholders' interests and needs?". Fairness at the policy level considers how the values of the organization building a system are translated into rules, leading to questions like: "does a policy prohibiting certain types of behavior within the product adequately address the unique experiences of some subpopulations?". Fairness at the implementation level deals with the empirical performance of the system, answering questions like: "Are human labelers executing the policy or labelling instructions correctly?" and "Are predictive models achieving the desired tradeoff between different types of errors for all subpopulations?". Notice that implementation fairness questions are not limited to machine learning models—we also study the human decisions used to train the models or directly intervene in the system.

These levels are of course deeply intertwined, but by separating and differentiating between them, we aim to direct analysis toward the most salient components of a system and determine whether and what changes are needed, while ensuring each component is appropriately considered in relation to the others. For example, in order to determine the relevant models within a product to assess for implementation concerns, practitioners must understand the intended goal, structure, and use of the product as well as any policies or similar rules that may have shaped or constrained the model's training data. If models are analyzed and found to have no implementation disparities across subgroups, but fairness concerns remain, practitioners then know to focus more deeply on the product design and policies in order to assess whether they appropriately consider the needs and harms of groups that may be impacted by that product. In other words, it is important to consider not only whether rules are being applied appropriately to all, but whether the rules themselves, or the structure in which they are situated, are fair, just and reasonable.

Unfairness stemming from a narrowly drafted policy can be escalated to and remedied by policy stakeholders, while unfairness resulting from poor machine learning implementation can be referred to machine learning engineers for remediation. Unfairness at the product level, meanwhile, may require a fundamental reimagination of a product's goals and objectives, requiring significant reallocation of resources by product leadership. By disentangling these three layers, unarticulated tradeoffs can be more explicitly enumerated, and disagreements about those tradeoffs can be situated with the relevant organizational decision-making frameworks.

We have found that disentangling fairness into component dimensions in this way more constructively facilitates conversations among those creating systems who are less familiar with the rich array of potential fairness-related harms and methodologies to address them, since technical work to investigate implementation concerns can be appropriately situated in qualitative conversations related to policies, legal obligations, user and stakeholder expectations, and real-world harm.

3.1 Fairness as a process

Fairness is an essentially contested concept [16], and significant interdisciplinary and intradisciplinary differences exist regarding how fairness is approached and evaluated. Scholars and practitioners from computer science, law, and philosophy, for example, may see fairness in very different lights, while related policy and regulatory notions may evolve over time. As such, there is often considerable disagreement about what fairness entails overall, let alone at a product-specific level. It is unrealistic to presume, and would be irresponsible to claim, that simply deploying tools, checklists, or frameworks is sufficient to fully mitigate fairness risks.

Indeed, a large proportion of fairness risks require weighing difficult tradeoffs, including seeking input from subject matter experts

[4] Jacobs and Wallach [27] note that the labels chosen to measure unobservable theoretical constructs should themselves be examined, but we place the choice of label beyond the scope of label bias for the purpose of this paper.

and people with lived experiences related to the potential harm, to inform the ultimate consensus that shapes how and to what extent fairness risks can be mitigated. Such deliberations require actionable ethical frameworks and qualitative research to fill in the gaps left by quantitative fairness approaches.

These processes must also be iterative in order to account for both evolving notions of fairness and justice, to allow for increasing degrees of sophistication in analyses, and to account for systems and products that themselves evolve over time. For example, subsequent analysis might broaden beyond a targeted assessment of an individual model to consider the more complex fairness questions that arise in compound or dynamic systems where multiple models may interact [15, 25], or ongoing monitoring may reveal that the cumulative effects of previously deployed fairness interventions have over time introduced unintended harms of their own, requiring the reevaluation of prior decisions about effective remedies to individual model-level unfairness.

Importantly, an iterative process acknowledges that fairness is never fully "solved," but rather encourages ongoing consideration of fairness throughout the product development lifecycle while preserving the ability to reassess what lens (or lenses) of fairness ought to apply in a particular context—and thus what method of assessment would be best suited to test for potential unfairness in that dimension. Creating space for such flexibility is an especially common need in areas where technology has illuminated a new, augmented, or resurgent fairness risk for which acceptable and expected remedies have yet to be defined, or for which consensus does not yet exist.

While an indispensable component of any holistic approach to fairness, we leave a detailed discussion of the opportunities and challenges of implementing such processes, as well as lessons learned from efforts to integrate them into organizational processes, to future work in order to discuss with sufficient degree of detail the challenges of addressing fairness at the implementation level in the context of a large and complex organization.

4 FAIRNESS IN IMPLEMENTATION

At its highest level of abstraction, our approach contains three steps: Catalog, choose, and monitor

(1) Catalog the costs and benefits the system may produce for people from different subgroups, and how these may trade off against one another.
(2) Make explicit choices about where the system should situate itself in this space of tradeoffs. Which costs and benefits should the system prioritize, and for whom? These normative decisions should be made at the product and policy levels of the fairness analysis.
(3) Ensure the system is minimizing costs and maximizing benefits according to the chosen tradeoffs in practice. This is the goal of the implementation level of the fairness analysis, which we address in detail in this section.

This approach most closely resembles one advocated for by Mullainathan [42], who argues that fairness analyses should proceed from a "description of a global welfare function." Kasy and Abebe [30] also study fairness in the context of costs and benefits; their approach to identifying the prioritization of groups (which they Broadly call the groups' "power") implied by observed decisions closely resembles what we discuss in Section 4.4.

While we refer to "costs" in the following sections, we use this term capaciously to describe not just economic costs but broader impacts of classification error, recognizing that such quantification requires a holistic understanding of potential harms to ensure they are sufficiently captured when weighing tradeoffs.

4.1 Benefits and harms in binary decision making

To see how this approach could be applied, consider a doctor deciding whether to prescribe chemotherapy to patients to treat potential cancers. The doctor cannot know for sure which patients have cancer, though they can divide the patients (perhaps using expert judgement, a diagnostic test, or a machine learning algorithm) into four categories of increasing risk, as shown in Fig. 1a. The probability of having cancer within each category is known, but there is no way for the doctor to identify exactly which patients within each category this probability will materialize for.

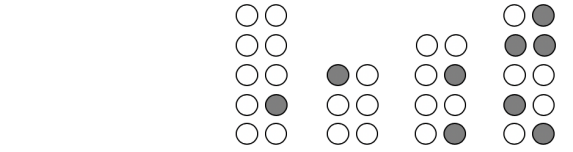
To a first approximation, there are two types of harms, or costs, to consider in this situation: the cost of a false positive, where the patient receives unnecessary and unpleasant chemotherapy; and the cost of a false negative, where the cancer goes untreated and may kill the patient.[5] There is a trade off between these costs—the aggressive prescription of chemotherapy will decrease false negatives but increase false positives, while a reluctance to pursue treatment will do the opposite. Resolving this tradeoff requires an assessment of, and judgement about, the *relative* cost of false positives and false negatives. Suppose that, after consulting with subject matter experts and relevant stakeholders, the doctor determines that an undiagnosed cancer is 4 times worse than unnecessary chemotherapy. This *cost ratio* captures what we mean by a system's tradeoffs—saying a false negative is four times more costly than a false positive is saying we would trade four false positives for one false negative, and vice versa. We note that in reality the choice to pursue a given medical treatment is also a highly personal decision that patients and their caretaking teams are involved in; we simplify the process here and assume that all patients have the same cost ratio for the purpose of illustration.

Fig. 1a shows the total cost of choosing to treat the patients in a given risk category (equal to the number of false positives this produces) compared to the total cost of choosing not to treat them (equal to four times the number of false negatives). Immediately, we can see the cost-minimizing strategy: treat only those patients in the two highest-risk categories. The optimal policy takes the form of a *threshold*—we treat everyone above a certain level of risk and no one below it. We can now think about the general case of choosing thresholds when patients each have a continuous risk score s, and false negatives are c times more costly than false positives (i.e. the cost ratio is c). Let

$$\text{cost}(t) = \text{FP}(t) + c\text{FN}(t) \tag{1}$$

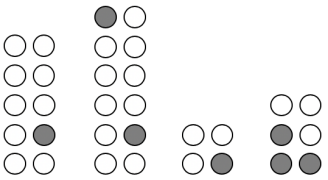
describe the total cost of all decisions when a threshold t is applied. We minimize the expected cost by taking the derivative with respect

[5] One could equivalently describe avoiding these costs as benefits, though this change of reference point wouldn't change the analysis



Probability of having cancer	10%	16%	25%	50%
False positives if treated	9	5	6	5
False negatives if untreated	1	1	2	5
Total cost if treated (1x FP)	9	5	6	5
Total cost if untreated (4x FN)	4	4	8	20
Cost-minimizing decision	Don't treat	Don't treat	Treat	Treat

(a) Female patients



Probability of having cancer	10%	16%	25%	50%
False positives if treated	9	10	3	3
False negatives if untreated	1	2	1	3
Total cost if treated (1x FP)	9	10	3	3
Total cost if untreated (4x FN)	4	8	4	12
Cost-minimizing decision	Don't treat	Don't treat	Treat	Treat

(b) Male patients

Figure 1: Hypothetical patients in a cancer treatment scenario. Dots symbolize patients at a given risk level, while solid dots denote patients who turn out to have cancer. Regardless of differences in the distribution of risk factors among the two different groups, the cost-minimizing treatment policy for both groups treats patients if and only if the probability they have cancer exceeds a certain threshold.

to t and setting it to zero:

$$\frac{d\,\mathbb{E}[\text{cost}(t)]}{dt} = \frac{d\,\mathbb{E}[\text{FP}(t)]}{dt} + c\frac{d\,\mathbb{E}[\text{FN}(t)]}{dt} \tag{2}$$

$$0 = -f(t^*)\,\mathbb{E}\left[1 - Y|s = t^*\right] + cf(t^*)\,\mathbb{E}\left[Y|s = t^*\right] \tag{3}$$

where $f(s)$ is the density of cases with a given score s and $\mathbb{E}[Y|s]$ is the rate of cancer cases among these patients. The derivative has a straightforward explanation: the decrease in false positives (or increase in false negatives) created by raising the threshold slightly is the product of the number of cases at the threshold and the fraction of those cases that were negative (or positive). Rearranging the equation gives us the optimality condition:

$$t^*_{\text{impl}} = \mathbb{E}\left[Y|s = t^*\right] = \frac{1}{1+c} \tag{4}$$

the cost-minimizing threshold is the score at which cases have a $1/(1+c)$ *probability of being positive.*[6] We call the probability $\mathbb{E}\left[Y|s = t\right]$ the *implied threshold*. Returning to our example we can now compute the precise cost-minimizing treatment strategy when $c = 4$: treat any patient with a greater than 20% chance of having cancer (i.e. the optimal implied threshold is 20%).

4.2 Introducing subgroups

Up until this point there has been no notion of subpopulations. Now imagine that the patients in Fig. 1a are female and the patients in Fig. 1b are male. This particular cancer affects a smaller fraction of male patients: 22% of males have it, compared to 26% of females. As a result, there are more males in the lower-risk categories and fewer in the higher-risk categories. We want ensure that our treatment strategy is fair to men and women.

[6]This threshold is a global optima as long as $\mathbb{E}[Y|s]$ monotonically increases in s, and it is the unique optimum if $\mathbb{E}[Y|s]$ is strictly monotonic. In practice it doesn't matter if the optimal threshold is not unique—there will be a range of thresholds (the closed interval between 1/6 and 1/4 in the example in Fig. 1a) that all produce the same, cost-minimizing decisions.

We have established the optimal treatment approach for female patients: treat all patients with a >20% chance of having cancer. But what is the optimal approach for male patients, who have a different base rate and distribution among the risk categories? Perhaps surprisingly, *as long as the cost of decisions is the same* (i.e. false negatives are four times more costly than false positives) the cost-minimizing approach is identical: we should treat all male patients with >20% chance of having cancer. This is because Eq. 4 does not depend on $f(s)$, the distribution of risk among the group in question; nor does it depend on the base rate.

This has important consequences for the study of fair machine learning. Consider an alternate approach that is popular in the fair machine learning literature: equalizing error rates [23]. One might argue that a false negative cancer diagnosis is so costly that fairness demands that our decisions produce equal false negative *rates* for male and female patients. The cost-minimizing approach does not produce such equality—the false negative rate for male patients is 43%, while the false negative rate for female patients is 22%.

Equalizing these rates would require some combination of treating more male patients and treating fewer female patients. However, since we've already chosen the treatment strategy that minimizes costs faced by both male and female patients, such an intervention would inevitably make at least one group worse off without making the other group better off. The same is true for equalizing false positive rates, or treatment rates (as would be required by demographic parity). A demand that these fairness criteria be satisfied, then, must reflect a judgement that equality in a particular ratio (in our example, the number of false negatives divided by the total number of cancer cases) is of greater fairness value than the minimization of the actual harm caused by errors to members of both groups. There are many reasons why this would not be an appropriate judgement to make in practice.

This is not to say that equal implied thresholds are always appropriate. Imagine if the cancer in question was more aggressive in female patients, such that a false negative was more likely to lead

to death. In this case, the cost of a false negative (relative to the cost of a false positive) is greater for female patients than male patients, and the cost-minimizing threshold for female patients decreases. Our approach to fairness, which focuses on concrete impacts to people, would therefore require that different thresholds be applied to treatment decisions for male and female patients. Note that this treatment regime, which is cost-minimizing for both groups, actually increases the differences between false negative rates for male and female patients. This further illustrates the disconnect between such error rates and the well-being of decision subjects.

4.3 Competing values

In cancer treatment, the potential cost of false positives *and* false negatives fall on the same patient, making it easy to argue for the decision-making strategy that makes patients from all groups best off [59].[7] But what about cases where the tradeoff occurs *between* groups? Consider the detection of spam on social media. The cost of a false positive principally falls on the publisher whose content is filtered (though consumers are also deprived of the opportunity to see this content), while the cost of a false negative falls on the consumer whose feed is contains low-quality content. As a result, the cost-minimizing approach for publishers would classify almost no content as spam,[8] while consumers would be better served by a more aggressive filtering system.

In these situations, a choice must be made about how the system will prioritize costs to different stakeholders. This prioritization doesn't have to reflect objective costs—cancer is objectively more costly than being caught in the rain, but it's perfectly acceptable for umbrella manufacturers to prioritize keeping people dry. Instead, it should reflect the goals and values of the system that have been carefully considered at the product and policy levels of fairness analysis. For example, we might decide that, for the spam filtering system, a false positive (affecting content producers) is ten times more costly than a false negative (affecting consumers). There is no single right answer here—different platforms make different tradeoffs in content moderation and different types of policy-violating content on the same platform might present significantly different costs—but being explicit helps ensure that the decision is made deliberately and consistently over time. The approach that achieves the desired prioritization is still described by Eq. 4: if false positives are tens times as costly as false negatives ($c = 0.1$), the optimal threshold is $^1/_{(1+c)} = 0.91$.

This is also true for any subset of the decisions. Consider the total costs created when potential spam published by a user from group a may be seen be a consumer from group b:

$$cost_{ab}(t_{ab}) = \text{FP}_{ab}(t_{ab}) + c_{ab}\text{FN}_{ab}(t_{ab}) \tag{5}$$

This takes the same form as Eq. 1, and as a result the cost-minimizing threshold depends only on c_{ab}, the cost to consumers from group b when they see spam from group a relative to the cost to publishers from group a when non-spam content is filtered out of a user from group b's feed. This motivates the key fairness principle that drives our analyses of binary decision making:

> *If we believe that a false positive is equally costly whether it affects somebody in group A or group B, and that a false negative is equally costly whether it affects somebody in group A or group B, we should apply the same implied threshold to decisions affecting both groups. Doing so will minimize errors no matter who they affect.*

The antecedent won't always be true; for example, a platform might decide to explicitly prioritize female publishers by treating a false positive affecting female publishers as more costly than one affecting male publishers. In the criminal justice domain, system designers might opt to treat false positives as more costly for groups who, for example, tend to be penalized more harshly in future circumstances for having a history of incarceration. Again, these are decisions that are most appropriately made at the product and policy levels of fairness analysis. If such a prioritization was agreed upon, our implementation fairness approach could still be used to ensure the appropriate—and in this case, different—implied thresholds are being applied in practice.

4.4 Analyzing existing systems

Many real-world systems are not designed according to the three-step approach laid out in Section 4. Rather than making explicit choices about how to trade off between different costs and benefits, they arrive at the set of decisions they make through heuristics, inertia, and—potentially—mistakes. In such cases, we may be called to assess the fairness of a system without knowing exactly how it was designed. Our approach can still be used in these circumstances, it just needs to be reversed. Instead of choosing a set of tradeoffs and then ensuring that they are achieved in practice, we can work backwards from the decisions the system is currently making to determine the set of tradeoffs (i.e. cost ratios) that make these decisions cost-minimizing. We call these the *implied* tradeoffs of a system.[9] In the case where the relevant costs are due to false positives and false negatives, we can invert Eq. 4 and use the implied threshold determine the cost ratio the decision maker is operating under:

$$c = \frac{1 - E[Y|s = t]}{E[Y|s = t]}. \tag{6}$$

Consider a treatment regime for the patients in Fig. 1 where female patients in the two highest-risk categories and male patients in the *three* highest-risk categories receive chemotherapy. Assuming this treatment regime minimizes some conception of cost for male and female patients, we can apply Eq. 4 to compute the implied cost ratios for both groups. $\mathbb{E}[Y|s = t, \text{female}] = 0.25$, since 25% of the patients at the threshold of treatment (the second-riskiest group) actually have cancer. This implies a female cost ratio of 3. In comparison, $\mathbb{E}[Y|s = t, \text{male}] = 0.16$, implying a male cost ratio of 5. This treatment regime implies we believe that false negatives are more costly for male patients than female patients (or, equivalently, that false positives are more costly for female patients). If this reflects our considered belief, then it is fine. If this was not

[7] Though, troublingly, some papers still advocate for decreased diagnostic performance in healthcare settings in the name of "fairness" [48].

[8] Scrupulous publishers might argue for some spam enforcement so that their good content doesn't get drowned in a sea of spam, but a concern for false positives affecting their content would keep them from advocating for a system as strict as consumers would prefer.

[9] Welfare economics and optimal taxation theory use a similar concept—"inverse welfare weights"—to describe how a decision maker must weigh different individuals' welfare make a set of decisions welfare-maximizing in the aggregate [30, 54].

Cancer treatment: FP and FN -> patient

Spam detection: FP -> publisher; FN -> consumer

intended, and instead occurred because of some error or oversight (for example, the systematic overestimation of male patients' cancer risk), then we can consider ways to fix this. Section 5 describes how we measure implied tradeoffs in algorithmic decision making and human labeling.

5 MEASURING FAIRNESS IN PRACTICE

To illustrate the principles behind our measurement approach, we'll use the example of a system designed to identify and remove posts that violate an online platform's policy against bullying and harassment. The system combines decisions made by humans with binary decisions made on the basis of machine-learned predictions. Human labelers are employed to determine whether a given post violates the policy; these labels then inform both immediate enforcement (i.e. the removal of posts from the platform) and the training of the machine learning system. The machine learning system, in turn, produces predictions that are used both to triage potential harassment for human review and to automatically remove the most obvious cases of bullying.

This example clearly illustrates the breadth of questions that must be answered before a technical fairness analysis is even attempted, and the utility of distinguishing among product, policy, and implementation fairness. For example, how should the policy define bullying and harassment? Do all expressions of harassment qualify, or does it depend on the subject of that bullying's membership in certain demographic groups? If the latter, which groups are protected? What is the nature and degree of harm potentially caused by false positives and false negatives, and how should the system trade off between those errors in enforcement? These are incredibly challenging questions (as evidenced by the fact that different internet platforms, social institutions, and liberal democracies take different approaches to objectionable speech) which we will not attempt to answer in this paper. Still, it is important to note that the empirical questions we consider at the implementation fairness level are only a small part of the fairness puzzle. worthy of objection

In this section we first apply our approach to decisions made by the machine learning system—the familiar problem of fairness in supervised learning. We then show that the same conceptual approach can be used to study the fairness of human decisions with respect to a ground truth.

5.1 Model fairness

The simplest and most ubiquitous type of algorithmic decision is a binary decision based on the prediction of a machine learning model—one action is taken if the probability of the predicted event is suitably high, another action taken otherwise. A system designed to automatically (without human intervention) identify and remove violating bullying and harassment takes this form.[10]

Consider the task of measuring the fairness of such a system. First, it is necessary to decide which subpopulations to compare. We could choose to measure the system for different groups of content consumers or producers, different groups targeted by bullying (which may not be the same as the intended consumer), or intersections of these groups. For the sake of exposition, though, let's consider two groups of content producers: groups "A" and "B". Differences in the base rate of outcomes are both extremely common in situations where fairness concerns are present (including in content moderation), and give rise to important impossibility results in the fair machine learning literature [33]. In our hypothetical scenario, group A's content is more likely to be bullying or harassment than group B's content.

Following the approach outlined in Section 4, we then have to identify the costs and benefits of the decision in question, and how these may be in tension for different groups. As in our previous examples, the most salient costs are the costs of false positives (erroneous removal of content) and false negatives (failing to remove bullying or harassment). There is no obvious tension between the costs experienced by *producers* in each group—taking a more lenient approach to content from group A, for example, does not require taking a more aggressive (or more lenient) approach to content from group B. Instead, the relevant tension is between the welfare of content producers and content consumers. To navigate this tension, we would need to be explicit about how we prioritize the removal of violating content from each group. For now, let's say we have decided that the appropriate cost ratio is the same for both groups: though group A produces more bullying content, an individual instance of bullying is no more costly coming from a user in group A than coming from a user in group B (we return to this decision in the next section). Similarly, a false positive is equally bad whether it affects users from group A or group B. As a result, our fairness principle in Section 4.3 holds, and we should expect to see equal thresholds being applied to content from both groups.

To ensure the desired tradeoff (equal cost ratios) is being achieved in practice, we must ensure the implied threshold is equal for content produced by both groups. It's important to distinguish between two different thresholds here. The *decision threshold* t defines how the output of the model (the score s) is mapped to decisions $\hat{Y}$ (i.e. whether the post is removed):

$$\hat{Y}_i = \mathbf{1}\{s_i \geq t\}, \tag{7}$$

whereas the *implied threshold* is the probability of the outcome corresponding to the decision threshold:

$$t_{\text{impl}} = \mathbb{E}[Y|s = t]. \tag{8}$$

If the score s is *calibrated*, such that $s = \mathbb{E}[Y|s = s]$, then the decision threshold and implied threshold are identical. In general, though, we cannot assume the model being assessed produces calibrated predictions (some models, like support vector machines, don't even produce scores in $[0, 1]$).

There are a number of challenges that must be overcome to estimate the implied threshold from the output of the actual machine learning system. First, when the scores are continuous the score will almost never take on exactly the threshold value. As a result, we could approximate the implied threshold by conditioning on the score taking some value within a window around the threshold:

First option $$\mathbb{E}[Y|s = t] \approx \mathbb{E}[Y|t_l \leq s \leq t_u] \quad \text{where} \quad t_l < t < t_u. \tag{9}$$

[10] Some might argue that such automated intervention would not be appropriate; such a conversation would be situated at the *product* level of our fairness analysis while investigations of the comparative performance of such a system would be a question of *implementation*.

Unfortunately, this approximation is often poor because of how scores are typically distributed. Bullying and harassment makes up a very small fraction of all posts on content platforms, so lower scores would generally be more common than higher scores, and samples in the window would disproportionately come from the lower end of the window. Therefore, simply averaging the outcomes of posts with scores within the window will tend to underestimate the implied threshold.

Alternative

To address these problems, we first fit a weighted linear regression to posts within a symmetric region (of half-width d) around the threshold using tricubic weights:

$$Y_i = \beta_0 + \beta_1(s_i - t) + \epsilon_i \tag{10}$$

$$w_i = \begin{cases} \left[1 - \left(\frac{|s_i - t|}{d}\right)^3\right]^3 & s_i \in [t-d, t+d] \\ 0 & \text{otherwise} \end{cases}. \tag{11}$$

Then, the intercept β_0 is our estimate of the implied threshold. We call this estimator the *prevalence at the threshold*. Statistically significant differences in this value between groups A and B would indicate that our system is not minimizing errors regardless of which group they affect, and is instead prioritizing errors affecting one group.

5.1.1 Alternative approaches. It is worth revisiting why we have not adopted alternative approaches to fair machine learning popular in the computer science literature. It's clear that statistical parity isn't appropriate: removing an equal fraction of posts from all groups, regardless of how much bullying they engage in, is not a viable approach for a real-world content moderation system. But the reason why we haven't chosen to, as a baseline, equalize error rates (either the false positive rate, the false negative rate, or both) is more subtle.

Existing research has established that the cost-minimizing approach to equalizing error rates uses different implied thresholds for each group [12]. In practice, the direction that thresholds must be adjusted is determined by a group's prevalence: equalizing error rates means applying a higher (more lenient) threshold to the group producing more bullying and harassment content (group A), and a lower (stricter) threshold to the group producing less of such content content (group B).[11] Since we know that the implied threshold and the cost ratio are related by Eq. 4, such differences in thresholds can only be rationalized by a decision to treat false negatives from group A (the higher-prevalence group) as *less* costly than the same errors affecting group B. In other words, equalizing error rates would imply that we believe that bullying and harassment is not as bad when produced by the higher-prevalence group.

It is important to note that there may be cases where false positives or false negatives *do* have different costs for content producers in group A or B, or for content consumers in group A or B; for example, if violating content produced by group B has a higher risk of leading to serious physical or psychological harms. In such cases, it may be deemed appropriate to apply different thresholds—but we hold that decisions about whether to treat subgroups differently as a fairness remedy should be made explicitly, with subject matter experts, and aligns with the *policy* or *product* dimension of fairness, rather than an unintended outcome of opting for statistical parity or equalized error rates within the *implementation* dimension.

Finally, error rates can depend on "easy" decisions that are irrelevant to the question of fairness. Most posts are obviously not bullying or harassment (e.g. "Happy birthday!"), and will never be removed by any content moderation system. And yet, since the number of non-violating posts makes up the denominator of the false positive rate, a group's false positive rate is affected by the number of obviously benign messages they post. It is clearly undesirable for a fairness assessment of a content moderation system to be affected by a group's tendency to share benign messages—but that would be the implication of considering false positive rates.

This problem is further exacerbated by the presence of adversarial behavior. Imagine bad actors from some group realizing that the system was designed to equalize false negative rates. They could spam the system with easy-to-identify instances of harassment, driving down their group's false negative rate. The system would be forced to respond by applying a more lenient threshold to content from the offending group, increasing the false negative rate to compensate and maintain error rate parity. By flooding the platform with obvious harassment, the bad actors would have forced us to be more lenient to the rest of their posts! This set of incentives would be problematic for a large-scale content moderation system to adopt. We note that as researchers continue to iterate on implementation fairness definitions and approaches, innovations in fair ML research may yet inform adaptations to our approach in the future.

5.2 Label fairness

As with most approaches to fair supervised learning, the approach described in the previous section assumes the outcome being predicted (Y) is measured accurately in the data used to assess the system. There are some cases where this assumption is reasonable—for example, websites can perfectly measure whether users click on a given button. However, in many cases, such as identifying bullying, the labels themselves are generated through human judgement, and may thus embed human biases. This is of concern for at least three reasons. First, accurate labels are needed to compute most model fairness metrics, including the metric in Section 5.1. Second, supervised learning systems trained on biased labels will learn those biases. Finally, labelers' decisions might be used to directly intervene in the system. In this section, we describe how our high level fairness approach can be applied to assess human decision making, in the case where decisions can be compared to a ground truth.

In our bullying and harassment example, the decision being made by human labelers is whether a given post violates a bullying policy as written. These decisions won't always be correct—labelers may misunderstand the policy or the post, make a mistake, or be misled by implicit or explicit biases. To track these errors, we also collect (for a subset of posts) the judgement of an expert in applying the written policy, whose decisions provide the ground truth for each

Another persona

[11] This is because, for most score distributions and thresholds, applying the same implied threshold to both groups will lead to a higher false positive rate and lower false negative rate for the higher-prevalence group. It is possible to construct score distributions where this does not occur (in which case equalizing error rates would require stricter thresholds for the higher-prevalence group), but these tend to be multi-modal in a way that we have not observed in practice.

post. A fairness measurement dataset, then, would consist of a set of tuples (Y_{ij}, Y_i^*) for every label, where Y_{ij} is the label provided by labeler j to post i, and Y_i^* is the expert-provided ground truth for that post.

As before, we can summarize the costs created by the human labeling process in terms of false positives and false negatives with respect to the ground truth. Ideally, then, we'd proceed as we would in the algorithmic decision making case: determine the implied threshold that labelers are using for posts from each group, and ensure it reflects the appropriate cost ratio. However, while in the algorithmic case we knew the decision threshold and which posts had scores close to the threshold, in the human decision making case we are missing all of this information. With only a binary label and a binary ground truth for each post, one might be tempted to abandon efforts to estimate the implied threshold and instead return to comparing something like group error rates. But this would be a mistake for the same reasons as described in Section 5.1.1: error rates are driven by easy decisions, both obviously benign (e.g. "Happy birthday!") and obviously violating. These are trivially easy for a human to judge correctly, and because of this are uninformative about a labeler's possible bias. We need a means of estimating the implied threshold with just a set of label/ground-truth pairs.

5.3 Signal Detection Theory

A promising approach comes in a model of human decision making developed by psychologists: Signal Detection Theory (SDT) [22]. Applied to the labeling of bullying or harassment, SDT models human decision making as follows: upon seeing a post, the labeler mentally accumulates evidence for and against the proposition that the post is policy-violating. The result of this accumulation is called the *signal*. The labeler also conceives of a threshold (sometimes called the *criterion*); when the signal exceeds the threshold they report the post as violating, when it doesn't they report the post as benign.[12]

Statistically, SDT models the distribution of the signal as a mixture of Gaussians: one Gaussian for benign posts and another for violating posts. Figure 2 illustrates the SDT model. Signal detection theory has been used to study human decision making in many different contexts since its development in the 1940s and '50s, including: military radar signals [22], medical diagnosis [58], child welfare decisions [43], and policing [49].

Since the scale of the decision variable is arbitrary, we can (without loss of generality) choose the distribution for benign posts to be the standard normal. By further assuming that the Gaussians have equal variance,[13] SDT defines a two parameter family of signal distributions parameterized by the *prevalence* (ϕ) and the *separation* (d', pronounced "dee-prime"). Prevalence is the fraction of posts that are violating, and therefore defines the mixing proportions of the Gaussians. Separation is the distance between the Gaussians' means. When separation is high there is little overlap between the signals produced by violating and benign posts, making it easy for labelers to correctly distinguish between them. When separation is low the distributions have substantial overlap, and labelers make more mistakes.[14]

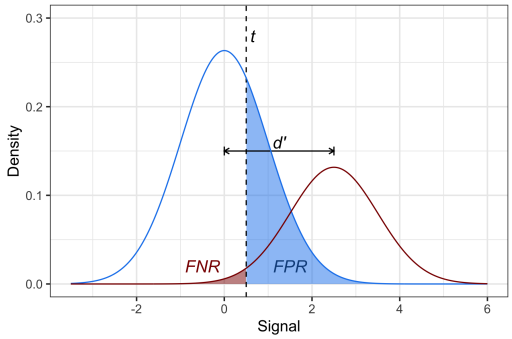


Figure 2: The signal detection theory model of decision making. Negative examples produce signals distributed according to the standard normal (blue curve). Positive examples produce signals sampled from a normal distribution with mean d' and unit variance (red curve). A threshold t is applied to turn signals into decisions. As a result, the false positive rate is simply the fraction of the negative distribution above the threshold (shaded blue), while the false negative rate is the fraction of the positive curve below the threshold (shaded red). Given observed false positive and false negative rates, we can therefore compute the t and d' values required to generate these error rates.

Figure 2 shows how, for a given separation and threshold, the signal detection theory model implies a certain false positive rate and false negative rate. In particular:

$$\text{FPR}(d', t) = 1 - \Phi(t)$$
$$\text{FNR}(d', t) = \Phi(t - d'),$$

where $\Phi(\cdot)$ is the cumulative distribution function for the standard normal. Therefore, we can use the observed error rates to infer the model parameters:

$$t = \Phi^{-1}(1 - \text{FPR})$$
$$d' = t - \Phi^{-1}(\text{FNR}).$$

Note that the threshold t is not the implied threshold, since it is defined in signal space. But SDT allows us to compute the implied threshold using the prevalence and Bayes' rule:

$$\begin{aligned} E[Y|s = t] &= P(Y = 1|s = t) \\ &= \frac{P(Y = 1)P(s = t|Y = 1)}{P(Y = 1)P(s = t|Y = 1) + P(Y = 0)P(s = t|Y = 0)} \\ &= \frac{\phi \text{N}(t - d')}{\phi \text{N}(t - d') + (1 - \phi)\text{N}(t)} \\ &= \frac{1}{1 + \frac{1-\phi}{\phi} \exp\left(-td' + d'^2/2\right)} \end{aligned}$$

[12] It's important to remember that SDT is a *model* of behavior—we're not suggesting that content labelers could report an actual numerical threshold if asked.

[13] Having equal variances ensures that the probability of a post being violating is monotonically increasing in the signal (see Pierson et al. [49] for proof and further details)

[14] In fact, separation is directly related to AUC, the probability that a randomly chosen violating post will have a higher signal than a randomly chosen benign post: $d' = \sqrt{2}\Phi^{-1}(\text{AUC})$ (where $\Phi^{-1}(\cdot)$ is the inverse cumulative distribution function for the standard normal).

($\mathrm{N}(\cdot)$ is the density function of the standard normal distribution.) Finally, we can plug the implied threshold into Eq. 6 to recover the cost ratio that explains labelers' decisions:

$$c = \frac{1-\phi}{\phi} \exp\left(-td' + d'^2/2\right). \quad (12)$$

Comparing cost ratios (or, equivalently, implied thresholds) between groups allows us to determine whether labelers are making appropriate judgements about content from different groups. In particular, if we believe labeling errors are equally costly regardless of the group they affect, we should ensure labelers are acting accordingly by applying the same implied threshold to all groups.

5.3.1 Limitations of the approach. While the two-parameter mixture model is flexible [49], it cannot capture all plausible signal distributions. For example, imagine there were three distinct types of posts: those that are obviously violating, those that are obviously benign, and those that are genuinely ambiguous. Signal detection theory cannot model such a tri-modal distribution, so it will produce incorrect estimates of the implied threshold in this case. Pierson et al. [49] attempt to address this by allowing the SDT parameters to vary according to decision covariates, but this requires a substantially more complex Bayesian inference procedure. We elect to use the simpler model to make it easier to scale the approach to many different labeling tasks.

An important direction for future work on labeling fairness centers on the separation parameter. A low d' for some set of decisions means that labelers have trouble distinguishing violating posts from benign posts. Currently, however, it is difficult to determine whether this trouble is due to the labeling problem being fundamentally hard in some sense, or because labelers are making mistakes—unconsciously or otherwise—in a way that may be attributable to their own bias (or other factors). For example, a labeler's bias against certain posts might manifest not in them erring towards false positives or false negatives (which SDT can measure using the cost ratio), but in them being indifferent about making errors in general, leading them to rush their decisions. In principle, this type of bias is still amenable to being measured with an extension of our tradeoff-focused approach—now the labeler is trading off between the cost of a false positive, the cost of a false negative, *and* the cost of their labeling time. The drift-diffusion model is an extension of SDT that attempts to account for decision time in this way [35].

6 PRACTICAL CHALLENGES AND OPEN QUESTIONS

6.1 Mitigating implementation fairness issues

Many machine learning papers that propose new model fairness metrics also develop algorithms to satisfy these metrics, either through optimization constraints [1, 12, 62] or by incorporating the metric into the training loss function [29]. These approaches purport to automatically ensure that a new model is "fair", but each necessarily reduces the performance of the model, increasing the number of people affected by model errors. We believe such approaches are often unwise: measured unfairness is a symptom of deeper problems in a system that likely can't be solved through a tweak in the optimization process. Furthermore, designing the optimization process such that fairness issues are never measured risks papering over these problems while often making decision subjects worse off.

Unfairness in a model has many different possible causes, including: a lack of training data, a lack of features, a misspecified target variable, or measurement error in the input features. None of these problems are amenable to typical machine learning optimization—their solutions exist outside the bounds of the optimization problem. The challenging upshot of this is that there is no silver bullet for mitigating implementation fairness issues. Instead, we believe that the measurement of fairness issues should prompt a deep dive into the model to diagnose and remedy the root cause of the issue.

This is especially true when trying to mitigate label bias concerns, since this always means changing human behavior. Fortunately, psychologists have demonstrated that labelers' will change the thresholds they apply when incentivized [13]. If the problem can be isolated to specific labelers who are being too strict or lenient, they could be be nudged into applying a more appropriate thresholds. If the problem is systematic, one should investigate the labelers' guidelines, how labelers are selected, and whether they are appropriately representative.

6.2 Group characteristic data

To measure potential bias affecting a sensitive subpopulations, one generally needs to know which people affected by the system are members of that group. However, the sensitive nature of the characteristics most relevant to fairness—including gender, ethnicity, religion, and national origin—poses important challenges for efforts to understand fairness in practice [9].

First, the entity trying to study fairness may lack subgroup information. While internet platforms may solicit a user's age and gender, they rarely collect or infer information about a user's race, ethnicity, religion, or sexual orientation. Collecting or otherwise obtaining this data raises privacy, ethical, and representational questions. In some cases, unresolved tensions between privacy and fairness have meant that we have lacked the data to perform fairness analyses pertaining to certain subgroups.

Even when the data is available, our methodologies require creating discrete groups out of complex identities. Discretizing someone's gender, age, or race will necessarily lack important details about their lived experience, or worse, may re-enforce historical categories that fuel discrimination or erase identities. But statistical measurement requires grouping users *somehow*. Fairness practitioners need to make sure that (a) group definitions and data are as reflective as possible of users' self-identities, (b) group designations are sufficiently flexible to capture a wide range of fairness concerns, and (c) users are provided sufficient control to fix mistakes in groupings.

Finally, subgroup information will always be subject to measurement error; even for self-reported attributes like gender users might decline to specify, choose an option at random, or make a data-entry mistake. Furthermore, some practitioners use *inferred* sensitive characteristics for fairness measurements and interventions [21]. Such inferences are likely to increase the number of errors in subgroup assignment by orders of magnitude. An open question remains as to how these errors could affect fairness measurements, especially if these subgroup identification errors are correlated with decision-making errors.

6.3 Complexity of systems

Using metrics that are correctly tailored to the potential benefits and harms that users may experience is central to our fairness approach. We have discussed metric recommendations for binary decision-making, but best practices do not yet exist to measure fairness for more complex model or system types.

For example, models are often combined or have feedback loops. In these systems, measuring only individual components may not reveal issues that emerge only in their interactions, and conversely, individual component measurements may not necessarily support drawing conclusions about an overall system. Conducting measurement on each single model is a necessary starting point, but further research into how components may combine to create—or reduce—fairness risk is needed.

7 CONCLUSION

This paper has presented an approach to addressing fairness challenges developed within the context of a large technology company. Our approach considers fairness at three levels—product, policy, and implementation—allowing us to direct analyses and interventions towards the appropriate part of the system, and to separate normative questions from statistical ones where appropriate. At the implementation level, we also presented a high-level approach to studying fairness questions grounded in the costs and benefits produced by decisions. Finally, we discussed that approach in two archetypal binary decision-making contexts: algorithmic decision making and human labeling.

REFERENCES

[1] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. 2018. A reductions approach to fair classification. *arXiv preprint arXiv:1803.02453* (2018).

[2] David Arnold, Will Dobbie, and Crystal S Yang. 2018. Racial bias in bail decisions. *The Quarterly Journal of Economics* 133, 4 (2018), 1885–1932.

[3] Gary S Becker. 2010. *The economics of discrimination*. University of Chicago press.

[4] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, et al. 2018. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943* (2018).

[5] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2018. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research* (2018), 0049124118782533.

[6] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Allison Woodruff, Christine Luu, Pierre Kreitmann, Jonathan Bischof, and Ed H Chi. 2019. Putting fairness principles into practice: Challenges, metrics, and improvements. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 453–459.

[7] Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. 2020. *Fairlearn: A toolkit for assessing and improving fairness in AI*. Technical Report MSR-TR-2020-32. Microsoft. https://www.microsoft.com/en-us/research/publication/fairlearn-a-toolkit-for-assessing-and-improving-fairness-in-ai/

[8] Miranda Bogen and Rieke Aaron. 2018. *Help wanted: An exploration of hiring algorithms, equity and bias*. Technical Report. Technical Report. Upturn. https://www.upturn. org/static/reports/2018 ....

[9] Miranda Bogen, Aaron Rieke, and Shazeda Ahmed. 2020. Awareness in Practice: Tensions in Access to Sensitive Attribute Data for Antidiscrimination. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 492–500. https://doi.org/10.1145/3351095.3372877

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

[11] Sam Corbett-Davies and Sharad Goel. 2018. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023* (2018).

[12] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*. 797–806.

[13] Tim Curran, Casey DeBuse, and P. Leynes. 2007. Conflict and criterion setting in recognition memory. *Journal of experimental psychology. Learning, memory, and cognition* 33 (02 2007), 2–17. https://doi.org/10.1037/0278-7393.33.1.2

[14] Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 67–73.

[15] Cynthia Dwork, Christina Ilvento, and Meena Jagadeesan. 2020. Individual Fairness in Pipelines. *arXiv preprint arXiv:2004.05167* (2020).

[16] Ronald Dworkin. 2002. *Sovereign virtue: The theory and practice of equality*. Harvard university press.

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

[18] Riccardo Fogliato, Max G'Sell, and Alexandra Chouldechova. 2020. Fairness Evaluation in Presence of Biased Noisy Labels. *arXiv preprint arXiv:2003.13808* (2020).

[19] Batya Friedman. 1996. Value-sensitive design. *interactions* 3, 6 (1996), 16–23.

[20] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2018. Datasheets for datasets. *arXiv preprint arXiv:1803.09010* (2018).

[21] Sahin Cem Geyik, Stuart Ambler, and Krishnaram Kenthapadi. 2019. Fairness-aware ranking in search & recommendation systems with application to linkedin talent search. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2221–2231.

[22] David Marvin Green, John A Swets, et al. 1966. *Signal detection theory and psychophysics*. Vol. 1. Wiley New York.

[23] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*. 3315–3323.

[24] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. 2019. Improving fairness in machine learning systems: What do industry practitioners need?. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–16.

[25] Lily Hu and Yiling Chen. 2018. A short-term intervention for long-term fairness in the labor market. In *Proceedings of the 2018 World Wide Web Conference*. 1389–1398.

[26] Lily Hu and Yiling Chen. 2020. Fair classification and social welfare. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 535–545.

[27] Abigail Z Jacobs and Hanna Wallach. 2019. Measurement and fairness. *arXiv preprint arXiv:1912.05511* (2019).

[28] James E Johndrow, Kristian Lum, et al. 2019. An algorithm for removing sensitive information: application to race-independent recidivism prediction. *The Annals of Applied Statistics* 13, 1 (2019), 189–220.

[29] Toshihiro Kamishima, Shotaro Akaho, and Jun Sakuma. 2011. Fairness-aware Learning through Regularization Approach. *Proceedings - IEEE International Conference on Data Mining, ICDM*, 643–650. https://doi.org/10.1109/ICDMW.2011.83

[30] Maximilian Kasy and Rediet Abebe. 2020. *Fairness, equality, and power in algorithmic decision making*. Technical Report. Working paper.

[31] Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. 2017. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems*. 656–666.

[32] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan. 2018. Algorithmic fairness. In *Aea papers and proceedings*, Vol. 108. 22–27.

[33] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807* (2016).

[34] John Knowles, Nicola Persico, and Petra Todd. 2001. Racial bias in motor vehicle searches: Theory and evidence. *Journal of Political Economy* 109, 1 (2001), 203–229.

[35] Ian Krajbich and Antonio Rangel. 2011. Multialternative drift-diffusion model predicts the relationship between visual fixations and choice in value-based decisions. *Proceedings of the National Academy of Sciences* (2011).

[36] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. In *Advances in neural information processing systems*. 4066–4076.

[37] Heidi Ledford. 2019. Millions of black people affected by racial bias in health-care algorithms. *Nature* 574, 7780 (2019), 608–609.

[38] Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. 2018. Delayed impact of fair machine learning. *arXiv preprint arXiv:1803.04383* (2018).

[39] Joshua R Loftus, Chris Russell, Matt J Kusner, and Ricardo Silva. 2018. Causal reasoning for algorithmic fairness. *arXiv preprint arXiv:1805.05859* (2018).

[40] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019.

Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*. 220–229.

[41] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. 2018. Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions. *arXiv preprint arXiv:1811.07867* (2018).

[42] Sendhil Mullainathan. 2018. Algorithmic fairness and the social welfare function. In *Proceedings of the 2018 ACM Conference on Economics and Computation*. 1–1.

[43] Jeryl L Mumpower and Gary H McClelland. 2014. A signal detection theory analysis of racial and ethnic disproportionality in the referral and substantiation process of the US child welfare services system. *1foldr Import 2019-10-08 Batch 3* (2014).

[44] Razieh Nabi and Ilya Shpitser. 2018. Fair inference on outcomes. In *Proceedings of the... AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence*, Vol. 2018. NIH Public Access, 1931.

[45] Helen Nissenbaum. 2001. How computer systems embody values. *Computer* 34, 3 (2001), 120–119.

[46] Ziad Obermeyer and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm that guides health decisions for 70 million people. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 89–89.

[47] Samir Passi and Solon Barocas. 2019. Problem Formulation and Fairness. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 39–48. https://doi.org/10.1145/3287560.3287567

[48] Stephen Pfohl, Ben Marafino, Adrien Coulet, Fatima Rodriguez, Latha Palaniappan, and Nigam H. Shah. 2019. Creating Fair Models of Atherosclerotic Cardiovascular Disease Risk. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (Honolulu, HI, USA) *(AIES '19)*. Association for Computing Machinery, New York, NY, USA, 271–278. https://doi.org/10.1145/3306618.3314278

[49] Emma Pierson, Sam Corbett-Davies, and Sharad Goel. 2018. Fast threshold tests for detecting discrimination. In *International Conference on Artificial Intelligence and Statistics*. 96–105.

[50] Ruchir Puri. 2018. Mitigating bias in AI models. *IBM Research Blog* (2018).

[51] Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. 2020. Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 469–481.

[52] Inioluwa Deborah Raji, Timnit Gebru, Margaret Mitchell, Joy Buolamwini, Joonseok Lee, and Emily Denton. 2020. Saving face: Investigating the ethical concerns of facial recognition auditing. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 145–151.

[53] Inioluwa Deborah Raji, Andrew Smart, Rebecca N White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 33–44.

[54] Emmanuel Saez and Stefanie Stantcheva. 2016. Generalized Social Marginal Welfare Weights for Optimal Tax Theory. *American Economic Review* 106, 1 (January 2016), 24–45. https://doi.org/10.1257/aer.20141362

[55] Javier Sánchez-Monedero, Lina Dencik, and Lilian Edwards. 2020. What does it mean to'solve'the problem of discrimination in hiring? social, technical and legal perspectives from the UK on automated hiring systems. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 458–468.

[56] Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 59–68.

[57] Camelia Simoiu, Sam Corbett-Davies, Sharad Goel, et al. 2017. The problem of infra-marginality in outcome tests for discrimination. *The Annals of Applied Statistics* 11, 3 (2017), 1193–1216.

[58] Thomas R Stewart and Jeryl L Mumpower. 2004. Detection and selection decisions in the practice of screening mammography. *Journal of Policy Analysis and Management* 23, 4 (2004), 908–920.

[59] Berk Ustun, Yang Liu, and David Parkes. 2019. Fairness without Harm: Decoupled Classifiers with Preference Guarantees *(Proceedings of Machine Learning Research, Vol. 97)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.). PMLR, Long Beach, California, USA, 6373–6382. http://proceedings.mlr.press/v97/ustun19a.html

[60] Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*. IEEE, 1–7.

[61] J. Wexler, M. Pushkarna, T. Bolukbasi, M. Wattenberg, F. Viégas, and J. Wilson. 2020. The What-If Tool: Interactive Probing of Machine Learning Models. *IEEE Transactions on Visualization and Computer Graphics* 26, 1 (2020), 56–65.

[62] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. 2017. Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. In *Proceedings of the 26th International Conference on World Wide Web* (Perth, Australia) *(WWW '17)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 1171–1180. https://doi.org/10.1145/3038912.3052660