Should Fairness be a Metric or a Model? A Model-based Framework for Assessing Bias in Machine Learning Pipelines

JOHN P. LALOR, AHMED ABBASI, and KEZIA OKETCH, University of Notre Dame, Notre Dame, USA
YI YANG, Hong Kong University of Science and Technology, Hong Kong, China
NICOLE FORSGREN, Microsoft Research, Seattle, USA

Fairness measurement is crucial for assessing algorithmic bias in various types of machine learning (ML) models, including ones used for search relevance, recommendation, personalization, talent analytics, and natural language processing. However, the fairness measurement paradigm is currently dominated by fairness metrics that examine disparities in allocation and/or prediction error as univariate key performance indicators (KPIs) for a protected attribute or group. Although important and effective in assessing ML bias in certain contexts such as recidivism, existing metrics don't work well in many real-world applications of ML characterized by imperfect models applied to an array of instances encompassing a multivariate mixture of protected attributes, that are part of a broader process pipeline. Consequently, the upstream representational harm quantified by existing metrics based on how the model represents protected groups doesn't necessarily relate to allocational harm in the application of such models in downstream policy/decision contexts. We propose FAIR-Frame, a model-based framework for parsimoniously modeling fairness across multiple protected attributes in regard to the representational and allocational harm associated with the upstream design/development and downstream usage of ML models. We evaluate the efficacy of our proposed framework on two testbeds pertaining to text classification using pretrained language models. The upstream testbeds encompass over fifty thousand documents associated with twenty-eight thousand users, seven protected attributes and five different classification tasks. The downstream testbeds span three policy outcomes and over 5.41 million total observations. Results in comparison with several existing metrics show that the upstream representational harm measures produced by FAIR-Frame and other metrics are significantly different from one another, and that FAIR-Frame's representational fairness measures have the highest percentage alignment and lowest error with allocational harm observed in downstream applications. Our findings have important implications for various ML contexts, including information retrieval, user modeling, digital platforms, and text classification, where responsible and trustworthy AI is becoming an imperative.

CCS Concepts: • **Computing methodologies** → **Machine learning approaches**; • **Social and professional topics** → **User characteristics;**

Additional Key Words and Phrases: Machine learning fairness, algorithmic bias, model framework, prediction and explanation, AI governance, machine learning pipelines

This work was funded in part through U.S. NSF grant IIS-2039915 and a Kemper Faculty Award.
Authors' addresses: J. P. Lalor, A. Abbasi, and K. Oketch, Mendoza College of Business, University of Notre Dame, Notre Dame, IN 46556, USA; e-mails: john.lalor@nd.edu, aabbasi@nd.edu, koketch@nd.edu; Y. Yang, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong, China; e-mail: imyiyang@ust.hk; N. Forsgren, Microsoft Research, Redmond, WA 98052, USA; e-mail: nicole.forsgren@microsoft.com.

ACM 1046-8188/2024/03-ART99
https://doi.org/10.1145/3641276

ACM Reference Format:
John P. Lalor, Ahmed Abbasi, Kezia Oketch, Yi Yang, and Nicole Forsgren. 2024. Should Fairness be a Metric or a Model? A Model-based Framework for Assessing Bias in Machine Learning Pipelines. *ACM Trans. Inf. Syst.* 42, 4, Article 99 (March 2024), 41 pages. https://doi.org/10.1145/3641276

1 INTRODUCTION

Machine learning (**ML**) models have become pervasive in every facet of our lives. They are widely used to automate or augment workflows and processes in real-world industry/organizational settings, and are the topic of exciting research spanning fields such as information retrieval, **natural language processing** (**NLP**), computer vision, customer analytics, and talent analytics. A great deal of predictive analytics research is geared toward improving the state-of-the-art of ML for user modeling [47, 48, 89], where models predict individual user-level outcomes, such as search relevance, recommendations, purchase propensity, engagement, and so forth [26, 34, 54, 64]. Although bias in computer systems is not a new issue, per se [42], the growth and omnipresence of ML models has ushered in a new set of concerns related to **artificial intelligence** (**AI**) governance, including concerns about algorithmic bias. Surveys of research on robust and trustworthy AI routinely talk about the need for fairness in ML [59].

While the literature on fairness and bias is vast, the overwhelming majority approaches measurement and mitigation in isolation (for example, refer to the many surveys on the topic that discuss the state-of-the-art [19, 34, 69, 78, 96]). For instance, the fairness of distributed representations of words is measured via tests such as WEAT [31], and debiasing is done with respect to these metrics. In addition, representations that make fairness guarantees have been proposed [24, 66, 67]. These methods have been shown to maintain fairness in the representations and also in predictive tasks for which the representations are used as inputs. However, the issue of whether these debiasing methods affect downstream models in a pipeline where the outputs of the initial task are used as inputs for subsequent decisions is often not considered. At the same time, *post-hoc* debiasing methods typically assume that inputs (e.g., learned embeddings) are fixed. In such contexts, fairness is measured using metrics such as disparate impact, and debiasing is done with respect to these metrics.

Recent empirical work has shown that there are relationships between biases (and debiasing methods) that occur *upstream* in an ML pipeline and those biases manifesting further *downstream* [45, 90]. For example, recent work has shown that debiasing word embeddings mitigates biases inherent in the learned embeddings but does not improve fairness when those debiased embeddings are applied to occupation prediction and toxic language classification tasks [90]. Generally speaking, this disconnect between upstream and downstream pieces of an ML pipeline can lead to disparities that are missed because they are not being measured in either location. In fact, prior empirical work has shown a lack of correlation between upstream and downstream fairness across multiple standard fairness metrics [45]. While, as mentioned, prior work typically focuses on either upstream or downstream biases, there have been recent empirical investigations into how the two may be linked. For example, the "bias transfer hypothesis" [23, 90] states that biases in **pretrained language models** (**PLMs**) such as BERT and GPT can be transferred to downstream models. If one were to zoom out and consider the data science environment more broadly, this transfer of bias could occur not only between a PLM and downstream model but also between intermediary ML models built in-house or accessed via a third-party **application programming interface** (**API**). As such, recent work has argued for a unified consideration of bias [109].

This empirical disparity between fairness assessment at different points in an organizational process has coincided with more theoretical investigations of *fair pipelines* [25]. These fair pipelines

concern decision environments where features are generated and/or decisions are made sequentially by multiple ML models. The theoretical work to date has focused on pipelines that filter out information at each step in the process (e.g., a hiring pipeline where candidates are screened out at each step) due to favorable theoretical properties. However, pipelines where estimated attributes and/or scores are passed through for all relevant individuals, are also widely used but not as widely studied in the literature. One can relate the steps in a pipeline to the notion of *representational* and *allocational* harms [19]. Recent work, in particular with regard to pretrained embeddings for text and/or images, has focused on the difference between representational and allocational harms. Allocational harms arise when models perpetuate an unfair distribution of resources. However, certain inputs to these downstream models may also generate representational harms. Consider a model such as BERT that is trained by a research group and made available online for download on the HuggingFace platform.[1] When a hiring firm downloads this model, representational biases in the model (e.g., the well-known "man is to programmer as woman is to homemaker" example [22]) can lead to allocational biases downstream. At present, the link between such biases is not well understood.

To be precise, we take the definitions of allocational and representational harms from prior work [19]:

> Allocational harms arise when an automated system allocates resources (e.g., credit) or opportunities (e.g., jobs) unfairly to different social groups; representational harms arise when a system (e.g., a search engine) represents some social groups in a less favorable light than others, demeans them, or fails to recognize their existence altogether.

Note in the above that representational harms can include both descriptive modeling (e.g., word embedding representations independent of a model) and predictive modeling that outputs a score, while allocational harms typically cover predictive modeling that leads to binarization (i.e., with some threshold), in order to assign scarce resources. In terms of pipelines, allocational harms occur at filtering steps, when individuals are denied access to opportunities further down the pipeline; representational harms occur at those steps in cumulative decision pipelines where ML model outputs are passed down to a later decision. In this work, we consider representational harms to be both the embedded stereotypes associated with canonical work [e.g., 22] and also those predictive modeling tasks such as sentiment analysis and psychometric trait measurement [9] where there is no direct allocation of resources, but harms can still be done (women are systematically scored as more anxious, men are scored as more negative, etc.). If upstream biases can be identified in a way that makes them both explanatory of the upstream bias and predictive of the downstream bias, then modelers and designers can incorporate this information into their pipeline-building process.

In the literature, there are many definitions and operationalizations of fairness. However, the aspirational goal of fairness has been defined as "the absence of any prejudice or favoritism toward an individual or group based on their inherent or acquired characteristics" [69]. As Barocas et al. [13] note (p. vi), "If machine learning is our way into studying institutional decision making, fairness is the moral lens through which we examine those decisions." Issues of fairness in ML can manifest at different phases of the model development process. Representational harm can occur when pretraining general-purpose language models or image embeddings. For instance, PLMs leveraging historical text corpora are known to capture stereotypes related to gender and race [15, 31, 44]. Similarly, image embeddings used in computer vision to train facial recognition models embody

[1] https://huggingface.co/models

gender and racial biases [27, 32]. Representational harm can also arise when fine-tuning models for specific tasks (i.e., fairness of upstream models). Commonly mentioned examples include the use of models to predict recidivism [38] or score applicants in talent recruitment contexts. In information retrieval contexts, the fairness of recommender systems has garnered considerable attention in recent years [64, 68, 96, 100]. User modeling work has also explored issues of fairness [47, 48], and there are implications for user heterogeneity and heterogeneous treatment effects [89]. Semantic representation bias associated with fine-tuned language models can also arise in various upstream modeling contexts [9, 36]. Moreover, as noted, these concerns are amplified by the increased usage of ML by data science and product teams in various industry practitioner settings [80].

Guided by the need to operationalize the tenets of responsible AI, there has been a significant uptick in computational research on fairness in ML, including taxonomies of fairness definitions and bias categories [13, 69, 96], research on developing and benchmarking debiasing strategies [46, 57], the proposal of new fairness measures [78, 103], examination of fairness in novel settings [107], and surveys on the current state-of-the-art and gaps [19]. In regard to the technical quantification and measurement of fairness, several important gaps persist [13]. Most existing research has focused either on measuring representational harm in upstream modeling contexts, or on the allocational harm associated with deploying such models in downstream application settings. The interplay between representational and allocational harm has received limited attention in the literature. Furthermore, existing computational research neglects the organizational data science environments in which ML model design, development, and deployment happen [25, 80]. Implementing new technologies to mitigate bias is not simply a technical problem for firms; it also involves stakeholder management, organizational strategy, and resource management, among other dimensions [80]. Stakeholders in these environments often approach ML modeling as an optimization problem involving a design search space [16, 88]. Moreover, how to handle protected attributes in a multivariate manner remains an open question.

Accordingly, the research objective of this study is to *develop and apply a model-based measurement framework for ML fairness that holistically measures upstream representational fairness and better predicts allocational fairness of downstream policy/decision outcomes in ML pipelines.* Our two research questions are as follows:

- **RQ1:** (a) How different are existing fairness metrics in their measurement of bias in ML representations? (b) Compared to fairness metrics, can fairness models better measure bias in ML across multiple protected attributes?
- **RQ2:** How accurately can fairness measurement of ML representations predict bias in downstream tasks/interventions?

To address these questions, we leverage the model typology literature, which notes the merits and strengths of descriptive, explanatory, and predictive models, and espouses more integration between explanation and prediction in relevant computational social science settings. We propose FAIR-Frame: **f**airness in **a**llocational, **i**nteractional, and **r**epresentational **frame**work. FAIR-Frame is a model-based framework for parsimoniously measuring fairness across multiple protected attributes in regard to the representational and allocational harm associated with the upstream design/development and downstream usage of ML models. The representational component of FAIR-Frame measures representational harm for a multivariate set of protected attributes across an ML model space with a search structure (e.g., a set of models derived across a grid or random search). The allocational component examines the extent of alignment between the upstream representational and downstream allocational measures. The interactional component takes into account the multivariate, interaction effect, linear/non-linear, and within versus between (ML model

design space) group facets of fairness measurement. In order to address our research questions, we evaluate the framework on two testbeds pertaining to text classification using PLMs. The upstream portion of these testbeds encompasses over fifty thousand documents associated with nearly twenty-eight thousand users, seven self-reported protected attributes, and five different classification tasks. The downstream testbeds span three policy outcomes and over 5 million total observations (each at the intersection of a user, task, upstream ML model, and downstream policy context). Results in comparison with several existing metrics show that the upstream representational harm measures produced by FAIR-Frame and other metrics are significantly different from one another (RQ1a), and that FAIR-Frame's fairness measures have relatively less variance and greater stability (RQ1b). Furthermore, FAIR-Frame also has the highest percentage alignment and lowest error with allocational harm observed in downstream applications (RQ2), yielding alignment rates that are 20-30 points higher than the best benchmark metrics.

The main contributions of this work are three-fold. First, we propose a framework for how fairness measurement may be able to move beyond univariate, descriptive metric-based measurement towards parsimonious, model inference-based assessment. We show that the current paradigm might be producing metrics with high variance and low parsimony that are only loosely associated with downstream allocational harm. These results are aligned with the notion that ML modeling and policy/decision spaces are multi-dimensional, multi-objective, and interactional, thereby necessitating simultaneous consideration of interactions and congruence with ML processes. Second, based on the model typology literature, our work underscores the value of integrated models of fairness that couple explanations in upstream representational contexts with the prediction of downstream allocational outcomes. Third, we apply our framework across two large testbeds spanning over 27 thousand users and 5.4 million downstream policy/outcome observations. To the best of our knowledge, no study has explored fairness measurement across multiple stages in such an in-depth manner in regard to the variables, tasks, models, and decision outcomes. Our work has implications for ML researchers exploring fair ML in important information retrieval, user modeling, and/or text/image analytics contexts, as well as practitioners interested in moving their workflows towards fairness-by-design principles [5]. Moreover, the proposed framework can be extended to other data modalities, ML modeling contexts, and real-world environments involving automation and augmentation of processes that implement "models on top of models" in ML/technology pipelines across an organization [6].

The remainder of this article is organized as follows. In the ensuing section, we discuss relevant prior work on fairness in ML, fairness metrics, and the model typology literature. In Section 3, we introduce our framework for measuring fairness in ML and conduct a theoretical analysis to demonstrate analytically how our framework addresses gaps in the existing literature. Section 4 describes our two testbeds for upstream and downstream fairness modeling, whereas Section 5 presents evaluation results comparing statistical differences in representational bias for our FAIR-Frame approach versus existing fairness metrics. Section 6 reports results for aligning representational harm metrics in upstream models with allocational disparities in downstream policy/decision settings. In Section 7, we conduct an initial analysis of the capability of FAIR-Frame to debias upstream predictions in an attempt to mitigate downstream allocation harms. In Section 8, we offer concluding remarks.

2 RELATED WORK

In this section, we review relevant prior work related to fairness in ML, fairness metrics, and the model typology literature. We conclude the section by highlighting the four major research gaps that underpin our proposed framework.

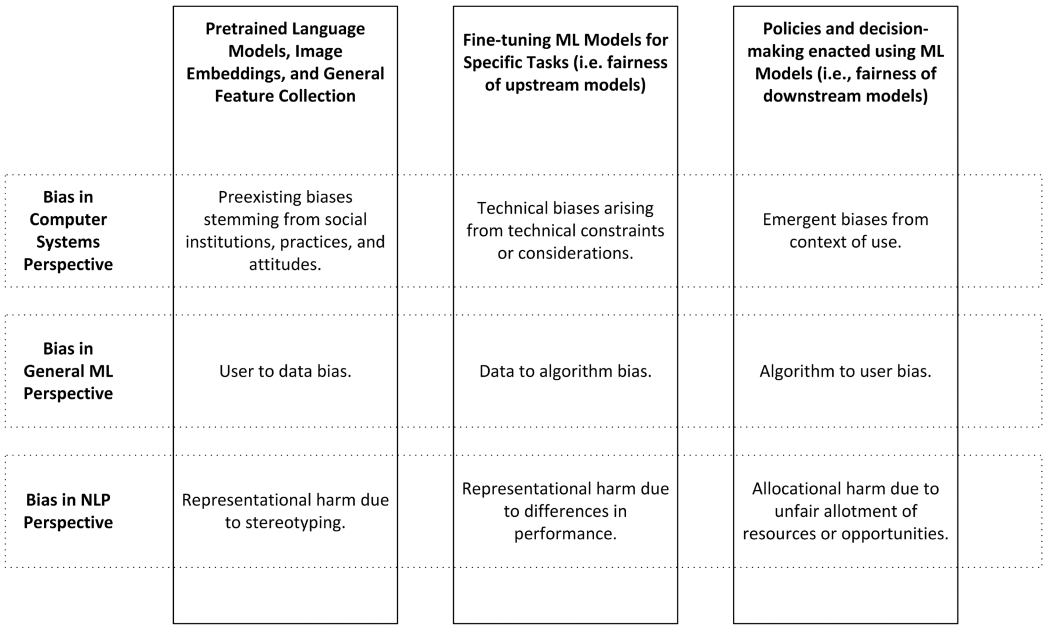


Fig. 1. Overview of fairness in ML across three stages from different perspectives.

2.1 Fairness in Machine Learning

2.1.1 Three Perspectives on Fairness. Algorithmic bias is the presence of unfairness in the data, models, and/or user experiences in contexts where ML is used to automate or augment processes or workflows, or where ML is used to enhance the user experience or value proposition. As noted earlier in the introduction section, fairness is the "absence of any prejudice or favoritism toward an individual or group based on their inherent or acquired characteristics" [69, p. 115]. Hence, fairness measurement in ML pertains to the quantification of ML biases across relevant phases of the ML modeling lifecycle. Rakova et al. [80] observed that the bias in computer systems literature dates back over 25 years [42]. More recent work has surveyed bias in ML algorithms [69], and bias in important classes of ML, such as NLP [19] and recommender systems [34, 96]. All three perspectives (computer systems, ML, and NLP/recommender systems) are important and consistent in describing three critical junctures of the ML modeling process where bias can manifest.

Figure 1 shows how the three perspectives (i.e., rows in the figure) describe biases in three phases of the ML modeling process (i.e., columns in the figure): the data used to pretrain; the fine-tuned (upstream) models; the (downstream) usage of the models. The first phase/juncture considers the data used to train ML models in general, text corpora to learn PLMs (for NLP), or image corpora to train large image embeddings (for computer vision). Such forms of bias have been described as preexisting biases attributable to institutions, practices, and/or attitudes that allow stereotypes to seep into the data and language models/image embeddings [13, 42]. In the context of PLMs, models trained on large historical corpora using techniques such as word2vec [70], GloVe [77], and contextual embeddings such as BERT [37] have been found to capture historical stereotypes related to gender and race [15, 31, 33, 44]. Similarly, image embeddings trained on large image corpora have also been found to embody gender and racial biases when used in contexts such as facial recognition [27, 32]. In recommender systems, historical item popularity affects which items are recommended in what's known as exposure bias [68].

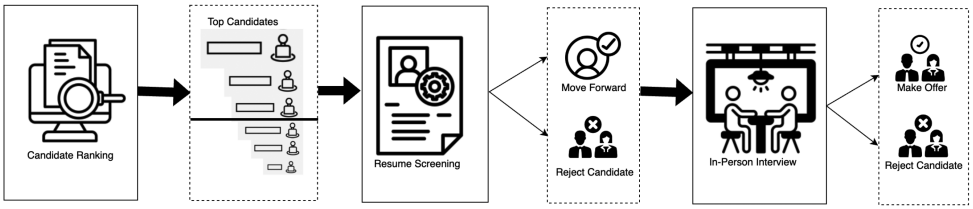
The second phase/juncture represents work that has focused on bias in models fine-tuned on the upstream data/tasks—often referred to as representational harm [13, 19], process fairness [96], or procedural justice [58, 72]. For instance, the recommender systems literature has explored the implications of user behavior biases on future model recommendations/performance for many years—namely, the effect of data/sampling biases pertaining to selection, position, exposure, and popularity [34, 64, 75, 108]. Some recent work has focused on the fairness of recommendation ML for parity in recommendation across item suppliers [68], two-sided/multi-stakeholder platforms [28, 100], and other system-induced biases [62]. Other recommendation fairness research has surveyed the role of gender and race biases in contexts such as job recommendations [96]. User modeling studies have also explored fairness of information retrieval ML [47, 48], the role of selection biases in question-answering systems [76], and bias in search rankings [83]. The text sequence classification literature has discussed the fairness of models for various applications, including psychometric NLP and text-based personality detection [9, 104].

The third phase/juncture represents work on policy- and decision-making as a result of the output of ML models or pipelines. This is where allocational harms arise when model outputs determine how and to whom benefits, decisions, job positions, and other such scarce resources are allocated. Trained models are typically used as aides in human decision-making processes. These models output probabilities that, when given a cutoff threshold, can assist decision-makers. This threshold can be determined by budget constraints (e.g., only a certain number of resources are available). Canonical examples here include the COMPAS tool for parole decisions [69, 78].

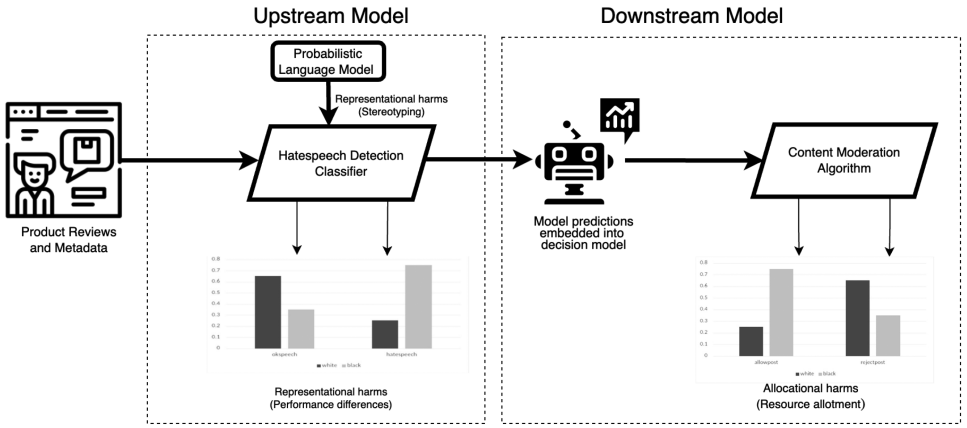
Looking at the current research, both on debiasing ML models and on fairness metrics, the focus has been on representational harm in pretrained models. Research across multiple stages that connect the outputs and implications of one stage with the inputs and consequences for a subsequent stage is lacking [19, 57, 85]. Additionally, managing fairness in ML in organizational and/or data science production contexts is a complex process involving many tradeoffs [71]. Responsible AI initiatives, such as bias measurement and mitigation, that do not consider the development process may lead to challenges and tensions between different stages [80]. For example, prior work has shown that state-of-the-art interventions from the research community to address fairness issues are often met with practitioner resistance, due to the difficulty of operationalizing new methods into existing and often mature technology systems [80]. While it is possible to implement end-to-end models that account for all three pillars, in the case of organizational deployment of ML models, firms typically take a modular approach to data science and model deployment, where teams train models for specific tasks and test and deploy them as part of a larger, interconnected pipeline.

2.1.2 Fair Pipelines. In the preceding discussion, the implicit assumption of the majority of fairness research is that ML models (e.g., models for computer vision, NLP, or recommendation) are considered in isolation. For measuring fairness, inputs and outputs are fixed, and for debiasing, either the inputs, model, or outputs are adjusted to optimize a predetermined fairness-aware auxiliary metric. Such models can be and often are learned in an end-to-end fashion. However, the landscape can also be considered as a sequential process (from left to right in Figure 1). There are scenarios where multiple ML models interact, mixing fixed inputs with estimated outputs through a pipeline to ultimately make an allocation decision. Mitigating biases in such systems have been studied under the moniker of *fair pipelines* [25].

Research on fair pipelines is emergent. Empirically, work has shown that biases in PLMs can be transferred to downstream models [23, 90]. Prior work has also shown a lack of correlation between upstream and downstream fairness metrics based on a wide range of available metrics [45], while other work has investigated a causal approach to identifying and mitigating biases in



(a) An example of a filtering pipeline, where there are machine learning models in the pipeline leading to an ultimate allocational decision of hiring a candidate.



(b) An example of a cumulative decision pipeline, where real-valued estimates from upstream ML models are used as input to a downstream allocational ML model.

Fig. 2. Two examples of pipelines that need to account for fairness differently.

PLMs used for downstream classification tasks [109]. Theoretically, recent research has defined fair pipelines and made theoretical guarantees about their fairness [25, 40, 65]. Prior research in fair pipelines distinguishes between pipelines with explicit allocation decisions at each step in the pipeline (*filtering pipelines*) and pipelines where features are estimated or scores are generated upstream and passed downstream to a later allocational ML model (*cumulative decision pipelines*) [25]. At present, filtering pipelines are more widely studied in the literature because allocation decisions binarize each step for probabilistic modeling [25, 40]. Figure 2 shows examples of filtering and cumulative decision pipelines.

Consider the hiring example of Figure 2(a), an oft-cited example in the fair pipelines literature. First, a recommender system for a hiring platform such as LinkedIn or Indeed ranks candidates for positions and assigns a relevance score. This score can either be used to restrict the pool of applicants to whom a job posting is shown (*filtering pipeline*) or as an input feature for firms downstream when assessing candidate quality (*cumulative decision pipeline*). Then, a (possibly automated) resume screening tool reduces the initial number of applicants. Finally, in-person interviews reduce the final candidate pool down to those candidates that the firm wishes to hire. At each step, fairness can be calculated based on the discrete, binary decisions of "move to next round" or "reject." For example, when hiring candidates, there are multiple allocation steps in the pipeline (reject or move forward at resume screening, reject or move forward at the phone screen, reject or make an offer after the in-person interview). While existing theoretical work has made

progress in understanding how and when fairness can be achieved in these pipelines, how these fairness calculations correlate across stages in the pipeline in terms of alignment and magnitude, and how multiple interacting demographics account for bias, is still unclear.

Filtering pipelines do not account for cases where ML models are embedded upstream *without* a subsequent allocational decision. To contrast the hiring example, consider an example from content moderation (Figure 2(b)). Suppose that a young, African American man logs onto an online shopping platform and writes a positive but somewhat lukewarm review for a product using the African-American English sociolect [AAE, 20]. The platform has a content moderation process in place that takes multiple pieces of information into account before deciding whether to accept or reject a posted review. Some of the input features are extracted directly from the customer and/or product metadata: customer demographics, product information, price, and so on. The firm also wants to predict whether the review contains potentially hateful speech so that they can flag these reviews for human inspection. For this, they outsource hatespeech detection to a third-party **prediction-as-a-service** (**PaaS**) provider [7]. The firm may not be aware that such models have been shown to encode certain biases with regard to AAE [20, 53, 56, 84]. Here, even though there is an upstream component (the hate speech scoring algorithm), the output is continuous, not binary. What's more, this algorithm does not filter out examples as in the hiring example but passes a score forward to the final allocation ML model.[2] While there is a growing body of work on fair pipelines, it is still not clear exactly how interventions manifest across stages. In fact, recent work suggested that "current notions of fairness compose poorly" [35]. This is especially true with regard to cumulative decision pipelines. The existing work on fairness binarizes the relevance score outputs based on a threshold in order to calculate an allocational fairness metric. Instead, what is needed is a notion of fairness at this representational level that can estimate the impact of demographics on the output score while also offering predictive power downstream.

Lastly, while prior work has made theoretical guarantees on the fairness of learned representations for classification decisions [e.g., 24, 66, 67], such claims do not extend to pipelines. As a brief, illustrative example, we train fair representations for a dataset of patient free-text reviews of medication. The method we use is from prior work on fair representations for user modeling [67]. Upstream, the task is to predict the sentiment of the text (e.g., positive or negative). This prediction can then be used as an input variable for a downstream prediction of whether or not the patient will continue to use the drug [8, 97]. We experimented with learning fair representations from **bag of words** (**BoW**) term frequencies and LIWC [93] to compare performance. Table 1 shows the difference in demographic parity between the protected ($\mathcal{Z}_1$) and privileged ($\mathcal{Z}_0$) groups: $\Delta_{DP}(g) \triangleq d_g(\mathcal{Z}_0, \mathcal{Z}_1) = |\mathbb{E}_{\mathcal{Z}_0}[g] - \mathbb{E}_{\mathcal{Z}_1}[g]|$. Although the initial upstream classifications are relatively fair, two issues emerge. First, the embeddings are made fair one demographic at a time.[3] Second, predictions in the downstream task are less fair and may vary in terms of which group the unfairness is directed toward. In Table 1, when moving from the upstream segment of the pipeline to downstream, the direction of the bias changes in 7 out of 8 scenarios.

Based on our overview, there are several things to note regarding fair pipelines. With regards to modeling, the PaaS model is upstream of the allocational model. In the example of Figure 2(b), it is not (and cannot be) trained end-to-end with the content moderation model because it is outside of the organization. There are two potential sources of representational harm here: the biased word embeddings which encode stereotyping of AAE text (Figure 1, Column 1), and the trained hatespeech detection model using the word embeddings such that input texts in AAE

[2]This step could be automated by setting an explicit score cutoff, which shows how pipelines can move between filtering and cumulative decision pipelines based on organizational decisions around automation.

[3]We note that certain algorithms consider multiple attributes but not at the subgroup level [e.g., 24, p. 722].

Table 1. An Example of Unexpected Behaviors of Fair Embeddings in a Pipeline Setting

		$\Delta_{DP}(g)$			
		Upstream		**Downstream**	
Features	**Debiasing**	**Gender**	**Age**	**Gender**	**Age**
BoW	Age	0.014	0.078	0.044	−0.306
	Gender	−0.115	0.186	0.060	−0.252
LIWC	Age	−0.015	0.033	0.039	−0.358
	Gender	−0.041	0.022	0.038	−0.379

The metric used to calculate fairness is difference in demographic parity, where a positive value indicates bias against the protected group and a negative value indicates bias against the privileged group. Unlike typical work, we do not take the absolute value so that we can indicate the directionality of unfairness. In this example, there are differences in alignment and in magnitude of the measured bias when plugging fair representations into downstream predictive tasks.

are systematically predicted as negative (Figure 1, Column 2). The internal content moderation model is downstream, and its decisions dictate whether there are allocational harms (e.g., allocating space on the webpage for the review, Figure 1, Column 3). It is in this multi-dimensional, interconnected landscape of ML models and fairness considerations within pipelines that we position our work.

2.2 Fairness Metrics

Having discussed *where* fairness issues manifest in ML models, we now discuss *how* such issues are measured. Fairness metrics can be categorized in a number of ways, according to how they are applied, what they are trying to measure, and what definition of fairness they are trying to achieve. For instance, in NLP, there is a large body of literature on the measurement of bias in PLMs using **word-embedding association tasks** (**WEAT**) and **sentence encoder association tasks** (**SEAT**) [31, 46, 57]. In this section, we discuss how AI models are evaluated with regard to the fairness of their *outputs*. Because most ML classification models output probability distributions over possible outcomes (e.g., if the probability that user-generated content posted to a social media site should be moderated is above some pre-defined threshold, then the model outputs a prediction of *Reject*). Therefore, metrics for assessing the fairness of these models use the probability distributions of the outputs to assess performance across different attributes.

The discussion below assumes that we have a binary classification task (e.g., content moderation), with a single demographic of interest that can be binarized into a *privileged* and *protected* split (e.g., male/female or young/old). That is, for some variable z, $\hat{z}$ is the (biased) estimate of z from an ML model, and d is the demographic dimension of interest, where $d = 1$ represents the protected class and $d \neq 1$ represents the privileged class. τ is some pre-defined threshold that acts as a cutoff to binarize $\hat{z}$. As we will discuss below, these metrics can be extended to the case of multiple demographics (i.e., intersectional fairness), but often have issues due to data sparsity when further segmenting the data.

Here we look at relevant work through the lens of Wang et al. [96] and distinguish between *Group Consistent* and *Calibrated* measures of fairness. Group consistent metrics make the implicit assumption that the treatment of two groups (i.e., privileged and protected demographic groups) should be consistent. These measures typically consider outcomes between two groups as a ratio with the expectation that the ratio be as close to 1 as possible. In addition, these metrics are calculated with respect to the model predictions but do not take into account base rates in the dataset used to train the model. Calibrated measures, on the other hand, look to balance the model outputs with the underlying base rates of the data distribution. These methods typically involve

Table 2. A Selection of Fairness Metrics from the Literature

Type	Metric	Formulation	References
Group Consistent	Disparate Impact	$\frac{P(\hat{z}\geq\tau\|d\neq 1)}{P(\hat{z}\geq\tau\|d=1)}$	[78]
	Demographic Parity	$\|P(\hat{z}\geq\tau\|d\neq 1)-P(\hat{z}\geq\tau\|d=1)\|$	[69, 78]
	Gini Coefficient	$\frac{\sum_{z_i,z_j\in Z}\|p(\hat{z}_i\geq\tau)-p(\hat{z}_j\geq\tau)\|}{2\|Z\|\sum_z p(\hat{z}\geq\tau)}$	[96]
Calibrated	Adjusted Disparate Impact	$\frac{P(\hat{z}\geq\tau\|d\neq 1)/P(z\geq\tau\|d\neq 1)}{P(\hat{z}\geq\tau\|d=1)/P(z\geq\tau\|d=1)}$	[57]
	Fairness Violation	$\|P(\hat{z}\geq\tau\|d\neq 1,z\geq\tau)-P(\hat{z}\geq\tau\|z\geq\tau)\|$	[91, 103]
	Equal Opportunity	$\|P(\hat{z}\geq\tau\|d\neq 1,z\geq\tau)-P(\hat{z}\geq\tau\|d=1,z\geq\tau)\|$	[69, 78]
	FPR Diff	$\|P(\hat{z}\geq\tau\|d\neq 1,z<\tau)-P(\hat{z}\geq\tau\|d=1,z<\tau)\|$	[69, 78]
	Jensen–Shannon Divergence	$\frac{1}{2}\left(\mathrm{KL}\left(\hat{Z}\|\|\frac{1}{2}(\hat{Z}+Z)\right)+\mathrm{KL}\left(Z\|\|\frac{1}{2}(\hat{Z}+Z)\right)\right)$	[96]

Note that Group Consistent metrics do not account for base rates, while Calibrated metrics do.

probabilistic outcomes conditioned, not only on the demographic class but also on the true outcome for a given example. In this way, the metrics are calibrated with respect to expectations built into the data, which may or may not have underlying biases in terms of how the data was collected (Column 1 in Figure 1). Table 2 lists the metrics and their formulations, which we describe in further detail below.

2.2.1 Group Consistent Metrics. The first set of metrics calculates fairness based on the model outputs, conditioned on the value of one or more demographic variables. Here, "group" signifies the demographic variable used to stratify the instances, and "consistent" refers to the aspirational goal of having similar predictions across strata. **Disparate impact** considers the ratio of positive predictions between protected and privileged groups [78]. The expected output should be close to 1 (e.g., the rate of positive prediction is similar between the two groups). In the legal literature, a threshold of 0.8 (or conversely, 1.2) is typically the allowable limit before the impact is considered to be disparate [14, 41]. **Demographic parity** (or statistical parity) takes the absolute difference between positive prediction rates, instead of the ratio [69, 78]. A score of 0 indicates a fair model, while values larger than zero indicate larger disparities. By construction, demographic parity does not consider the direction of the disparity, only its presence. **Gini Coefficient** is a metric commonly used in the recommender system literature. It comes from the field of economics and is widely used to measure a degree of individual unfairness in a distribution (e.g., income levels in a region) [96]. Lower values indicate more fair predictions. Gini Coefficient also only measures magnitude and not direction of unfairness.

2.2.2 Calibrated Metrics. Calibrated metrics are similar to group consistent metrics, but include the additional constraint of conditioning on true outcome values in order to account for base rates. Here, "calibrated" refers to the notion that we want to consider predictions and true labels, with the latter signifying base rates. **Adjusted disparate impact** extends disparate impact by additionally normalizing according to base rates to normalize the positive prediction rates [57]. As with disparate impact, values close to 1 are considered fair. **Fairness violation** examines how different the positive predictions for individuals in a protected class vary from the rate of positive

predictions for the dataset overall [91, 103]. The expectation here is that this difference is close to 0. If the value is negative, then positive cases are underpredicted for the protected group compared to the overall population. If the value is positive, then positive cases are underpredicted for the privileged group. **Equal opportunity** looks at the absolute difference between the **true positive rate** (**TPR**) for the privileged subset of the data and the TPR for the protected subset of the data [69, 78]. As with demographic parity, 0 indicates a fair model, while values larger than zero indicate larger disparities. Here again, the direction of the disparity is not measured. **False positive rate difference** is similar to equal opportunity but considers false positive rates instead of TPRs [69, 78]. **Jensen–Shannon Divergence** is a distance metric to measure the difference between two probability distributions [96]. In the context of fairness, it measures the distance between the distribution of predicted outcomes and the distribution of true outcomes. A smaller value indicates a fairer model.

In all of the above cases, the most common approach is to calculate fairness with respect to a single demographic attribute (e.g., race or gender). In the case of *intersectional fairness*, where more than one demographic group is under consideration [24, 101], data sparsity can affect the results [57]. Conditioning on multiple demographics reduces the space of examples considered for each measure, which can lead to high variance metrics across combinations of demographics. What's more, it is not clear how different interactions are related to each other. Instead, production data scientists and researchers must inspect a collection of metrics that are independently calculated to assess some general notion of fairness for candidate models. What is needed is a more parsimonious way to calculate and inspect fairness, not only across demographics (and combinations of demographics) but also in a holistic way that considers magnitude and direction of the potential bias.

2.3 The Model Typology Literature

The model typology literature has noted that there are three main types of models: descriptive, explanatory, and predictive [86, 87, 105]. Descriptive modeling describes situations in the past or present [50], but is neither causal nor predictive. Descriptive models do not provide causal or correlational effect sizes, may not include constructs, and generally lack sound theoretical basis for, and perhaps even the presence of, relations between independent and dependent variables [50, 86]. Conversely, explanatory models estimate the effect sizes of independent variables on a dependent variable of interest, typically using association-based regression-style models [50, 86, 105]. Similarly, predictive modeling is "the process of applying a statistical model or data mining algorithm to data for the purpose of predicting new or future observations," [86]. Predictive models are geared towards "out of sample prediction" [50, 79]. Explanatory and predictive models consider bias-variance tradeoffs; however, the former focuses more on bias reduction to ensure models best fit the data and theoretical constructs, whereas the latter (predictive) models consider variance more due to the implications for expected prediction error [86, 105]. More recent work has advocated for a fourth class of models that combine elements of explanatory and predictive models for contexts where explanation and prediction might be equally beneficial [50].

The model typology literature is relevant to the ML fairness measurement context in several ways. First, most existing fairness measures (e.g., those described in Table 2) are descriptive in nature. They do not parsimoniously model the relations between protected attributes and model performance/error. Second, integrated models that explain bias in upstream ML contexts and predict it in downstream policy/decision settings have been underexplored, despite the fact that such models could be advantageous for aligning measurement of representational and allocational harm. We expound upon this gap in our subsequent research gaps section and propose a mechanism for addressing this gap as part of our framework discussed in Section 3.

2.4 Research Gaps

Based on our review of prior work on fairness in ML, fairness metrics, and the model typology literature, we tackle the following research gaps that are important to fill in order to better model and understand fairness in ML in upstream and downstream contexts. These gaps are closely aligned with the two research questions explicated in the introduction section and filling these gaps provides an important contribution to the literature discussed previously:

— *Examining Downstream Allocational Harm* - As noted in the Introduction and Related Work sections, the majority of work on fairness in ML has focused either on representational harm in pretrained models or representational/allocational harms in fine-tuned models [e.g., 34, 64, 75, 108]. Fairness research is fragmented between the models and their applications [17] and does not account for ML-generated data passed downstream to another ML model in a pipeline. There is a need for research that normatively examines allocational harm in downstream settings [19, 57]. This is important for a myriad of reasons. Downstream allocational harm is related to, but different, from upstream representational harm—understanding both is crucial to gauge the broader AI governance implications of ML models more holistically. What's more, this understanding in feature extraction pipelines, where allocational decisions are made at the final step, has yet to be studied. For instance, downstream usage contexts of the ML model may include process augmentation, where the model outputs serve as an input to a human or model-based decision—with the hope that the use of ML can improve decision quality, agility, and operational outcomes [54, 102].

— *Congruence with ML Modeling Process* - Fairness work often neglects the organizational and/or project team environments in which ML models are designed, developed, and deployed [59, 80]. First, existing work often considers the fairness of a single model in isolation, whereas they are typically designed and deployed as part of pipelines within organizations. For instance, "how effects of bias compound in decision-making pipelines" involving multiple models has been underexplored [25]. Second, ML projects involve searching over a solution space [16], comprising a set of model configurations, predictive task formulations, and/or debiasing strategies to meet multi-variate organizational requirements. Applied data scientists and researchers typically use mechanisms, such as random/grid search, AutoML, **neural architecture search** (**NAS**), model ablation analysis [74], and so on to navigate this space [39, 49, 81]. Realizing the aspirational goal of fairness by design necessitates connecting fairness measurement with the (upstream) model design and (downstream) deployment environments [5].

— *Modeling Multivariate Interaction Effects* - Prior work in fairness has taken a univariate perspective of a protected attribute [30, 57, 92]. This approach makes it difficult to effectively compare multiple protected attributes at once, or demographic interactions between attributes (often called intersectional bias), resulting in unstable, high-variance measures. Balancing fairness and predictive power metrics can be challenging, even for a single protected attribute, depending on the overlap (or lack thereof) of the conditioned **receiver operating characteristic** (**ROC**) curves for different sub-groups [12, 94]. These challenges get exacerbated in interactional contexts. For instance, prior work has shown that interaction effects can cause fairness metric range and variance to increase considerably, making it difficult to disentangle algorithmic bias from fairness measurement limitations and relate protected class attributes to predictions and outcomes [57].

— *Towards Parsimonious Integrated Models* - Most prior fairness work has proposed descriptive metrics or models that are not capable of offering parsimonious explanations of fairness in upstream models, or predicting downstream fairness [94]. Integrated models that can

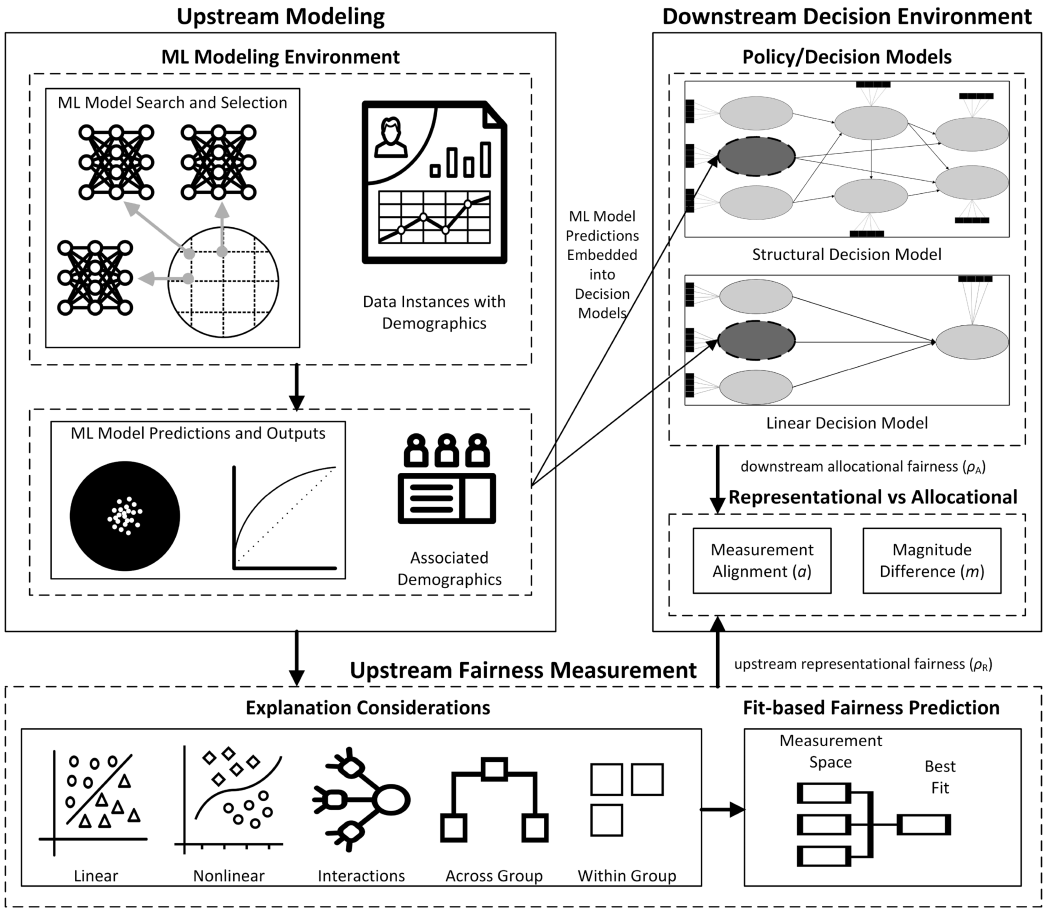


Fig. 3. Overview of fairness framework that considers upstream representational and downstream allocational harm (FAIR-Frame).

explain upstream representational harm and correlate with, or predict, downstream allocational harm could be valuable and more actionable in practical real-world contexts.

3 PROPOSED FAIRNESS MODELING FRAMEWORK

3.1 Framework Overview

In this work, we propose FAIR-Frame, a framework for modeling fairness that parsimoniously models the magnitude and direction of bias. FAIR-Frame seeks to both explain representational fairness through a parsimonious model of feature estimation error and also predict allocational fairness when the estimated features are used as part of a downstream policy decision task. Figure 3 presents FAIR-Frame.

The top left panel depicts the (upstream) ML modeling environment that uses a grid/random search to develop a set of viable models. Notably, the model development environment is not the focus of our work. The outputs of this module are the model predictions and associated instance-level demographics. These outputs are input into (downstream) linear or structured policy/decision models (right panel in Figure 3). The inclusion of ML outputs in the downstream models is visually

signified by the darker-colored ovals. Relative to the use of a gold-standard measure, this inclusion introduces prediction error which could produce allocational harm attributable to the use of imperfect ML approximations. The bottom panel denotes the multivariate models used to measure representational harm (again using the output of the upstream modeling environment). The use of models allows us to consider linear/non-linear patterns, across/within/nested structures for the ML models, and interactional biases. FAIR-Frame can also select the best-fitting fairness models based on established statistical measures. As shown in the bottom part of the right panel, we can measure the alignment and error between the representational harm (bottom panel) and allocational harm (right panel) approximations. In the remainder of the section, we describe the framework in greater detail.

3.2 The FAIR-Frame Method

As a motivating example, consider the moderation of user-generated content online shown above in Figure 2(b) [1, 29]. The decision of whether to flag content for moderation may include several variables, such as a prior history of the user posting the content, the web page to which the content was posted, and flagged keywords that appear in the content. User profile information and demographics may also be included if they are known. An ML model can also analyze the content itself, e.g., to determine the sentiment of the text, whether the text includes hateful speech, or psychometric characteristics of the individual who provided the text [1, 4, 9, 43]. The hatespeech detection model output can be used as part of the downstream moderation decision model. There are two estimates in this picture, the (possibly biased) estimate of the upstream feature (hatespeech detection), and the (possibly biased) estimate of the downstream policy decision (moderation).

While such tasks can be done in an end-to-end fashion, it is becoming more and more prevalent for firms to leverage PaaS providers to extract predictions from data [7]. PaaS providers offer a high-performance alternative to expensive data collection and in-house model training/storage. In this example, the firm could use Microsoft Azure's Content Filtering[4] or Amazon's Toxicity Detection services.[5]

Ideally, representational bias is in alignment with the allocational bias. This way we have a predictive element in our representational bias calculation that can inform how the downstream model will perform with regard to demographic subsets of the data [50]. In the rest of this section, we provide a detailed description of FAIR-Frame. Table 3 describes the key variables for reference.

3.2.1 Upstream Modeling. For the upstream case (e.g., the hate speech score of a user-generated review), we have some variable z; we want to train a model to predict z, $\hat{z}$. For this task, we have training data where the gold-standard values for z are known. To predict z requires models trained on some relevant dataset (e.g., a hate speech detection dataset). There is a large space of possible models that can be selected to estimate the variable of interest, including models that are trained and/or fine-tuned with fairness as a constraint. Considerations for how a model is selected may include accuracy, interpretability, costs associated with training and storage space, number of parameters, and so forth. In a real-world scenario, developers use tools and/or heuristics to search the possible solution space for candidate models [16, 39, 49, 74, 81]. For some space of models Φ and a training dataset X, the set of models trained on a training dataset X is the selection space for the upstream modeling task:

$$\Phi^* = \{\phi(X) | \phi \in \Phi\}. \tag{1}$$

[4] https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/content-filter
[5] https://aws.amazon.com/transcribe/toxicity-detection/

Table 3. Descriptions of Variables Used in FAIR-Frame

Variable	Description	Equation Reference
Φ	ML model space	1
X	Training dataset	1
Φ^*	Upstream model selection space	1
ϕ^*	Trained model from selection space	2
z	True upstream output variable	2
$\hat{z}$	Model estimate of z	2
d	Demographic information	2
Δ_R	Representational error	2
ρ_R	Representational fairness	2
π	Downstream policy decision	3
$\hat{\pi}$	Downstream policy decision prediction	3
γ	Downstream policy model	3
Δ_A	Allocational error	6
ρ_A	Allocational fairness	6
a	Fairness alignment	7
m	Fairness magnitude difference	8
$\hat{\Delta}_R$	Estimated representational error	9
$\mathcal{M}$	FAIR-Frame model space	

Each trained model $\phi^* \in \Phi^*$ can be evaluated by generating predictions on a benchmark dataset X_B. This benchmark dataset must include the input data (user-generated content), output label (contains hate speech), and demographic characteristics of the individual who created the input. The demographic information may or may not be used during training, but is required for calculating fairness. Training with demographic features is considered *fairness through awareness*, while training without the demographic features is known as *fairness through unawareness* [69].

The output of model ϕ^* on all examples in X_B, $\hat{\mathbf{z}}$, is a biased estimate of $\mathbf{z}$; we can, therefore, calculate the error. We call this error *representational error* (Δ_R) as it relates to the error in the representation of our feature z. What's more, because this feature was learned from a dataset that includes demographic information, we can also calculate the *representational fairness* (ρ_R) of this feature with respect to demographics [13, 19]:

$$\begin{aligned} \hat{\mathbf{z}} &= \phi^*(X_B), \\ \Delta_R &= \mathcal{L}(\hat{\mathbf{z}}, \mathbf{z}), \\ \rho_R &= F_R(\hat{\mathbf{z}}, \mathbf{z}, \mathbf{d}, \tau_R), \end{aligned} \tag{2}$$

where $\mathbf{d}$ is the vector of demographic variables for each example in the dataset, indicating whether a particular instance is a member of the protected group or not, F_R is a predetermined fairness metric (e.g., those in Table 2), and τ_R is a predefined threshold to binarize $\hat{\mathbf{z}}$ and $\mathbf{z}$ for group fairness calculations. The loss function $\mathcal{L}$ is flexible enough to handle regression or classification tasks. However, $\mathcal{L}$ must be such that its range includes positive and negative values. For example, a squared loss would not be appropriate, because then we would not be able to tell between overprediction and underprediction for a protected class (i.e., all loss values would be positive). Our framework can include both *fairness through awareness* (i.e., $d \in X$) or *fairness through unawareness* (i.e., $d \notin X$). In both cases, $d \in X_B$ because it is needed to calculate ρ_R. The estimated output from ϕ^*, $\hat{\mathbf{z}}$, is now used by the downstream policy model as an input.

3.2.2 Downstream Decision Environment. We assume that there is some downstream decision to be made that involves allocation [13, 19]. For example, in the case of user-generated content,

the decision may be whether to allow the user's post to stay up or to remove it. The decision itself (allow/withdraw content) is π. Let γ be some predictive *policy model* that is tasked with providing a prediction for π, $\hat{\pi}$, based on some inputs Z, where $\hat{z}$ is a subset of Z and includes one or more of the input parameters, as estimated by the upstream ML model:

$$\begin{aligned} \hat{\boldsymbol{\pi}} &= \gamma(Z), \\ \hat{\mathbf{z}} &\in \mathbf{Z}. \end{aligned} \tag{3}$$

We assume that the policy model must have some considerations for explainability [82] (a linear model, structured model, etc.). The linear model consists of some linear combination of input features and may or may not include a non-linear transformation (NL):

$$\hat{\boldsymbol{\pi}} = NL(\beta_0 + \beta_d * \mathbf{Z} + \epsilon). \tag{4}$$

A structured downstream representation [21] is more complex. Here we assume that the model is defined by a series of equations that specify latent endogenous ($\boldsymbol{\eta}$) and exogenous ($\boldsymbol{\xi}$) variables, observed variables ($\mathbf{Z}$, $\boldsymbol{\pi}$), intercept terms ($\boldsymbol{\alpha}$), learned coefficients ($\mathbf{B}$, $\boldsymbol{\Gamma}$, $\boldsymbol{\Delta}$), and error terms ($\boldsymbol{\zeta}$, $\boldsymbol{\epsilon}$, $\boldsymbol{\delta}$):

$$\begin{aligned} \boldsymbol{\eta}_i &= \boldsymbol{\alpha}_\eta + \mathbf{B}\boldsymbol{\eta}_i + \boldsymbol{\Gamma}\boldsymbol{\xi}_i + \boldsymbol{\zeta}_i, \\ Z &= \boldsymbol{\alpha}_x + \boldsymbol{\Lambda}_x\boldsymbol{\xi}_i + \boldsymbol{\delta}_i, \\ \hat{\pi} &= \boldsymbol{\alpha}_y + \boldsymbol{\Lambda}_y\boldsymbol{\eta}_i + \boldsymbol{\epsilon}_i. \end{aligned} \tag{5}$$

In this work, we assume that those models in $\mathcal{M}$ have been identified and constructed in a rigorous way. It is not our aim to optimize prediction or fairness with regard to the policy models. FAIR-Frame is intended to *model* the process of existing models based on human decisions and model estimation procedures that are consistent with best practices [21].

For some set of inputs, the model γ will have some degree of error compared to the gold standard policy decision. We refer to this error as *allocational error* (Δ_{A}) as it refers to the degree to which incorrect decisions affect the allocation of downstream resources (i.e., approved or denied posts). Along with allocational error, we must also consider how the model performance varies across some protected demographic attribute or attributes (e.g., race or gender):

$$\begin{aligned} \Delta_{\mathrm{A}} &= \mathcal{L}(\hat{\boldsymbol{\pi}}, \boldsymbol{\pi}), \\ \rho_{\mathrm{A}} &= F_{\mathrm{A}}(\hat{\boldsymbol{\pi}}, \boldsymbol{\pi}, \mathbf{d}, \tau_{\mathrm{A}}). \end{aligned} \tag{6}$$

We call this *allocational fairness* (ρ_{A}), where $\mathbf{d}$ is the demographic characteristic(s) under consideration for the fairness calculation [13, 19] and F_{A} is a fairness metric (Table 2).

3.2.3 Metrics. Based on the fairness metrics defined above (ρ_{R} and ρ_{A}), we construct the following metrics to evaluate the predictive performance of representational fairness metrics. For concreteness, both metrics must be *consistent* in how they are calculating fairness. Without loss of generality, consider the following: if a negative value for ρ_{R} represents representational harm against the protected class, then a negative value for ρ_{A} must also represent allocational harm for the protected class. For example, an upstream Fairness Violation score that is negative (Table 2) means that positive predictions for the protected class, conditioned on being a true positive, are lower than positive predictions across classes, conditioned on being a true positive. If we expect biases to propagate through a model, then the downstream fairness violation would also be negative. In this context, high **alignment** indicates the predictive power of representational fairness for allocational fairness, the latter being the more crucial final decision [25]. Alignment calculates whether the directionality of the metric output is consistent between the representational and allocational fairness calculations. If there is high alignment, then both fairness metrics are capturing

consistent information with respect to which group is unfairly treated.

$$a = \mathbb{I}[\text{sgn}(\rho_A) = \text{sgn}(\rho_R)]. \tag{7}$$

Going back to the previous example, a relatively small negative fairness violation score upstream would indicate that although there is bias against the protected class upstream, it is minor, and we would expect that to be consistent downstream. A deviation from this (e.g., a large negative fairness violation score downstream) would mean that there is an amplification of the bias when the upstream variables are plugged into the downstream model. This motivates our second metric, **magnitude difference**. Magnitude difference is an assessment of how accurate the representational metric is in predicting the value of the downstream metric.

$$m = |\rho_A - \rho_R|. \tag{8}$$

3.2.4 Representational Fairness Measurement. We propose that, instead of treating representational fairness (ρ_R) as a metric to be measured, we treat it as a model to be estimated. We hypothesize that modeling representational fairness will lead to better calibration (i.e., lower magnitude difference and higher alignment) with allocational fairness (ρ_A) than descriptive metrics such as those currently used in the literature. To do this, we introduce a procedure for modeling ρ_R by predicting Δ_R. That is, we formulate representational fairness as a model to predict representational error as a function of demographic attributes. This model can include demographic interactions to account for intersectional allocational fairness as well.

Let $\mathcal{M}$ be the space of parsimonious models for estimating Δ_R. We first separate the input data X into demographic variables X_d and other variables X_c. Models in $\mathcal{M}$ can be structured in a number of ways, accounting for linearity/non-linearity, across- and/or within-group separations, and interactions, among others. These models all take the following form:

$$\begin{aligned} \theta &= \beta_0 + \beta_d * X_d + \beta_c * X_c + \beta_i * (X_d)^{\psi} + \epsilon \qquad \textit{Within} \\ \theta &= \beta_0 + \beta_d * X_d + \beta_c * X_c + \beta_i * (X_d \times \Phi^*)^{\psi} + \epsilon \quad \textit{Between} \\ \hat{\Delta}_R &= \sigma(\theta). \end{aligned} \tag{9}$$

Here, ψ indicates the degree to which parameter interactions are estimated by the model. $(X_d)^{\psi}$ represents interactions within our demographic characteristics to model intersectional biases. When $\psi = 0$, there are no interactions modeled. σ can either be some non-linearity applied to the linear combination of weights and inputs, or it can be an identity transformation. Equation (9) gives a flexible and interpretable framework for estimating Δ_R as a function of demographics. Moreover, this formulation is aligned with prior work seeking high-performing and interpretable models [82]. Parameters for FAIR-Frame are learned by minimizing the loss between the true error and the estimated error:

$$\arg\min_{\beta} L_{\text{FAIR-Frame}}(\Delta_R, \hat{\Delta}_R). \tag{10}$$

For each model in the space, parameters are estimated. Performance characteristics of the estimated models are used to determine the estimation model to calculate fairness. Each model $m \in \mathcal{M}$ has an associated measure of performance based on number of parameters, likelihood, and so on. We select the estimated model parameters associated with the best-performing model as our learned estimates of fairness. In our experiments below, we use model fit, specifically, the **Akaike information criterion** (**AIC**) to determine the best model for fairness [10]. AIC penalizes models with more parameters; when comparing models, a lower AIC score is preferred.[6]

[6] AIC is a function of the number of model parameters k and the maximum likelihood estimation of the model L: $2k - 2\ln(\hat{L})$.

Table 4. Descriptions of Additional Variables Used in Our Theoretical Analysis

Variable	Description	Equation Reference
$\mathcal{F}$	Pipeline decision functions	11
$\mathcal{D}$	Pipeline decision outputs	11
$\mathcal{G}$	Pipeline rule functions	12
$Pipe(\mathcal{F}, \mathcal{G})$	Pipeline of decision and rule functions	13
$\hat{y}_t$	ML model output at pipeline step t	16
U^*	Pipeline utility	16
α_t	Budget constraint at pipeline step t	16

The learned coefficients (β) represent the effect of each demographic variable on the representational error of the upstream model, Δ_R. A positive coefficient indicates *underprediction* for the protected class of a given demographic. That is, $\Delta_R = z - \hat{z}$ is larger, indicating that the predictions are lower than the true values. On the other hand, if the coefficient is negative, then the model is *overpredicting* z for the protected class of the demographic. In the context of the user-generated online content example, if $\beta_{\text{Age}} > 0$, then predicted hate speech scores for texts written by older individuals are lower than the true values, indicating underprediction for older individuals. A fair model would have demographic coefficients equal to 0 (or not significantly different from 0), so that demographic information does not have a significant effect on representational error.

To summarize, the FAIR-Frame framework models representational fairness instead of measuring it in order to better explain the contribution of multivariate inputs to unfair outputs. At the same time, modeling fairness is more predictive of downstream allocational fairness [50], as we show in our theoretical analysis below.

3.3 Theoretical Analysis

We present a theoretical analysis to underscore some of the main research gaps, and how FAIR-Frame offers positive affordances related to these gaps. Namely, based on the fair pipelines literature, we highlight the importance of downstream fairness and of modeling fairness across links in ML pipelines. We illustrate the utility cost of measuring and enforcing fairness upstream. Lastly, we shed light on the biased estimates produced by univariate metrics. Newly introduced notation is summarized in Table 4.

3.3.1 Preliminaries. Following prior work, we define a pipeline as follows [25]:

Definition 3.1 (Pipeline). An n-stage pipeline $Pipe(\mathcal{F}, \mathcal{G})$ on set O is defined as an ordered set of decision functions $\mathcal{F}$ and rule functions $\mathcal{G}$:

$$\mathcal{F} = \{f_1 : O \to D_1, f_2 : O \times \widehat{D_1} \to D_2, \ldots, f_T : O \times \widehat{D_{T-1}} \to D_T\}, \tag{11}$$

$$\mathcal{G} = \{g_1 : D_1 \to \{0,1\}, \ldots, g_{T-1} : D_{T-1} \to \{0,1\}\}. \tag{12}$$

Here $\widehat{D_t} = D_{k_t} \times D_{k_{t+1}} \times \cdots \times D_{t-1} \times D_t$ for $1 \le k_t \le t$. Thus, the final decision for $x \in O$ is the output of the last decision function, which takes into account all prior information:

$$Pipe(f, g)(x) := \begin{cases} \hat{f}_T(x) & \text{if } g_t(\hat{f}_t(x)) = 1 \ \forall t \in [T-1] \\ \text{FAIL} & \text{otherwise} \end{cases}. \tag{13}$$

Here $\hat{f}_t(x) = f_t(x)$ for $t = 1$; $\hat{f}_t(x) = f_t(x, \widehat{f_{k_t}}(x), \widehat{f_{k_{t+1}}}(x), \ldots, \widehat{f_t}(x))$ for $2 \le t \le T$ and $1 \le k_t \le t$.

Remark. Pipelines can also account for additional input features at a given f_t; for simplicity, they are not shown here [40]. This formulation of pipelines can account for filtering and cumulative decision pipelines. Applying filtering pipelines according to this framework requires the existence of a suitable τ_t at each stage such that $g_t = \hat{f}_t(x) \geq \tau_t$. Note that stage i is *upstream* of stage j if $i < j$.

Definition 3.2 (Local Fairness). Local fairness is satisfied if, at each stage in the pipeline, the specified fairness metric holds. For example, if the metric under consideration is demographic parity (Table 2), then in a locally fair pipeline the following holds for each t, $1 \leq t \leq T$:

$$P(\hat{y}_t = 1 | \hat{y}_{t-1} = 1, d = 0) = P(\hat{y}_t = 1 | \hat{y}_{t-1} = 1, d = 1). \tag{14}$$

Definition 3.3 (Global Fairness). Global fairness is satisfied if, at the final stage in the pipeline, the specified fairness metric holds. For example, if the metric under consideration is demographic parity (Table 2), in a globally fair pipeline the following holds:

$$P(\hat{y}_T = 1 | d = 0) = P(\hat{y}_T = 1 | d = 1). \tag{15}$$

Prior work has shown that if a pipeline satisfies local fairness, then it implies global fairness [40].

Lemma 3.4. *A locally fair pipeline (i.e., one that is fair at each point in the pipeline) is globally fair.*

Definition 3.5 (Pipeline Utility). The utility of a filtering pipeline U is defined as a sequence of predictions $(\hat{y}_1, \ldots, \hat{y}_T)$ that maximize precision (or some other performance metric) while respecting fairness and budget constraints.[7] Budget constraints are defined by the percentage of entries allowed to pass through in a given stage: $(\alpha_1, \ldots, \alpha_T)$ where $\alpha_t = 1 - \tau_t$ and τ_t is the binarization threshold discussed earlier (Section 2.2).

$$\begin{aligned} U^*(\boldsymbol{\alpha}_{T-1}, \alpha_T) = & \max_{\hat{y}_1 \ldots \hat{y}_T} P(y_T = 1 | \hat{y}_T = 1) \\ & \text{subject to} \\ & P(\hat{y}_t = 1) \leq \alpha_t, \; t \leq T-1 \\ & P(\hat{y}_T = 1) = \alpha_T \\ & \rho_j(\hat{y}_1, \ldots, \hat{y}_T) = 0, j \leq T, \end{aligned} \tag{16}$$

where ρ_j is the fairness metric at step j in the pipeline.

3.3.2 The Case Against Measuring Upstream Fairness: Price of Local Fairness. At present, measuring upstream fairness in pipelines implies that a threshold is imposed, so that empirical probabilities can be calculated. However, many pipelines do not have upstream thresholds (e.g., cumulative decision pipelines). What's more, the tradeoff for local fairness is decreased overall utility, as defined by a linear program optimized with prediction and fairness as constraints [40]. We first show that measuring and enforcing local fairness leads to decreased overall utility in the pipeline.

We can calculate the ratio of the utility of a globally fair model (U_G^*) that has a single fairness constraint ρ_T and a locally fair model (U_L^*) with multiple fairness constraints $(\rho_1, \ldots, \rho_T)$ to determine whether the utility is gained or lost by choosing one method over the other.

[7] Depending on the position in a pipeline, the prediction y_t can be a thresholded representational output (e.g., $\hat{z} \geq \tau_t$) or allocational output (e.g., $\hat{\pi}$).

Theorem 3.6. *A filtering pipeline that is fair at each point suffers from the Price of Local Fairness [PoLF, 40].*

$$PoLF = \frac{U_G^*(\boldsymbol{\alpha}_{T-1}, \alpha_T)}{U_L^*(\boldsymbol{\alpha}_{T-1}, \alpha_T)}, \tag{17}$$

$$1 \leq PoLF(\boldsymbol{\alpha}_{T-1}, \alpha_T) \leq \min(\frac{1}{\alpha_T}, \frac{1}{P(\hat{y}_T = 1)}). \tag{18}$$

Proof. Due to the fact that a locally fair algorithm imposes more constraints than a globally fair algorithm, $U_L^*(\boldsymbol{\alpha}_{T-1}, \alpha_T) \leq U_G^*(\boldsymbol{\alpha}_{k-T}, \alpha_T)$ and therefore $1 \leq PoLF$ [40]. On the other hand, if there are no budget constraints (e.g., $\alpha_t = 1 \; \forall t < T$), then the utility is determined by the lesser of the final filter α_T and the probability of a true label ($P(\hat{y}_T = 1)$) [40]. □

Remark. Imposing constraints upstream (both in terms of budget and in terms of additional fairness metrics) in a pipeline has lower utility (higher PoLF) than simply calculating global fairness. What's more, if the final allocation decision involves the allocation of rare resources (e.g., $P(\hat{y}_T = 1)$ is very small), then the PoLF becomes substantial. The PoLF result suggests that you should not threshold upstream when calculating fairness if you do not need to. This is particularly relevant for cumulative decision pipelines, where there is not an upstream filtering decision. By imposing an artificial filter upstream to fit existing fairness measurements, and then (presumably) debiasing to improve the performance, we are imposing a cost on the global fairness of the pipeline. One should instead defer fairness measurement to the allocational decision, and model fairness upstream.

3.3.3 Bias in Unidimensional Fairness Measurement. In a pipeline, we expect that upstream estimations of fairness are predictive of downstream fairness so that organizations can set expectations with regards to the fairness of the downstream allocational decisions made by the ML model. In this section, we show that a metric for upstream fairness that only looks at a single demographic (as most existing metrics in the literature are designed to do), when used as a predictor of downstream fairness, is unpredictably biased compared to a model-based approach that considers all relevant demographics in a single model.

Theorem 3.7. *Discrete measures of upstream fairness are biased as predictors of downstream fairness.*

Proof. Consider a model to predict downstream fairness as a function of upstream fairness. The upstream fairness score could be either the output of an upstream metric (e.g., ADI) or the estimated parameters from FAIR-Frame. Assume without loss of generality that we have a two-stage pipeline and two demographic variables associated with gender and race, respectively: d_1 and d_2. For typical measure-based metrics, where each upstream fairness value is measured separately (e.g., ρ_{R} is the upstream fairness for gender, calculated independently from considerations of race) we have the following model to estimate downstream fairness:

$$\tilde{\rho}_{\mathrm{A}} = \tilde{\zeta}_0 + \rho_{\mathrm{R}} d_1 + \mu. \tag{19}$$

Here, the tilde indicates that our estimate is from the underspecified model. For FAIR-Frame, we have the fully specified model of downstream fairness ($\hat{\rho}_{\mathrm{A}}$) which includes coefficients for gender and race:

$$\hat{\rho}_{\mathrm{A}} = \hat{\zeta}_0 + \beta_1 d_1 + \beta_2 d_2 + \mu. \tag{20}$$

We can estimate the effect of β_2 on ρ_{R} via a simple regression and define ρ_{R} in terms of β_1 and β_2 by setting the two regressions to be equal, as they are both estimating the same quantity, ρ_{A} [99]:

$$d_2 = \tilde{\delta}_1 d_1, \tag{21}$$

Table 5. Testbed Overview

Location	*Testbed Characteristic*	*Description*	
		Psychometric	AskAPatient
Upstream	Number of Users	8,395	18,186
	Number of Documents	33,580	18,186
	Demographic Variables	Age, Race, Gender, Income, Education	Age, Gender
	ML Model variables	Anxiety, Trust, Numeracy, Literacy	Drug Rating
	ML Models	BERT, RoBERTa	BERT, RoBERTa
	Fairness Adjustments	Debiasing Yes/No	Debiasing Yes/No
Downstream	Policy variables	Well Being, Doctor Visits	Drug Duration
	Policy models	Structured, Linear	Linear
	Policy Modeling Decisions	Feature interactions, non-linearity	Feature interactions, non-linearity
	Total number of observations	5,372,800	36,372

$$\tilde{\zeta}_0 + \rho_R d_1 + \mu = \hat{\zeta}_0 + \beta_1 d_1 + \beta_2 d_2 + \mu, \tag{22}$$

$$\rho_R d_1 = \beta_1 d_1 + \beta_2 \tilde{\delta}_1 d_1, \tag{23}$$

$$\rho_R = \beta_1 + \beta_2 \tilde{\delta}_1. \tag{24}$$

We now calculate the expected value of ρ_R to determine the bias:

$$\mathbb{E}(\rho_R) = \mathbb{E}(\beta_1 + \beta_2 \tilde{\delta}_1), \tag{25}$$

$$= \mathbb{E}(\beta_1) + \mathbb{E}(\beta_2)\tilde{\delta}_1, \tag{26}$$

$$= \beta_1 + \beta_2 \tilde{\delta}_1, \tag{27}$$

$$\text{Bias}(\rho_R) = \beta_2 \tilde{\delta}_1. \tag{28}$$

□

Remark. Here note that not only is the model biased, but it is *unpredictably* biased [99]. Depending on the sign of β_2 and the correlation between d_1 and d_2, the bias could be positive or negative. So we cannot know *a priori* whether our upstream metric will be positively or negatively correlated with the downstream metric, which affects our prediction ability and also our estimation of the magnitude of the bias. In practice, typical measures of group fairness tend to be positively biased [65]. We also note here that prior work has shown that applying off-the-shelf fairness metrics for intersectional biases where multiple attributes are considered together exacerbates the perceived magnitude of the bias [57].

4 TESTBED OVERVIEW

To evaluate FAIR-Frame, we constructed a testbed that encompasses upstream ML variable modeling as well as downstream policy decision making. Table 5 describes the data for our experiments, which we discuss below.

4.1 Psychometrics

The **Psychometric** dataset includes free-text responses that are linked to four psychometric measures: Health Literacy, Health Numeracy, Trust in Doctors, and Anxiety Visiting Doctors [3]. Users were given psychometric tests in these four areas. These survey responses were collected and

scored according to standard psychometric techniques [73]. The survey respondents were also asked to answer a free-text prompt to describe their feelings on the dimension of interest. For example, for the construct of Anxiety Visiting Doctors, the prompt was: "In a few sentences, please describe what makes you most anxious or worried visiting the doctor's office"[3].

This dataset has been used in prior work to estimate the psychometric score from the free text using NLP models [3, 57]. In our context, we assume that the psychometric property is an input feature to the policy decision, and is estimated by the upstream ML model (i.e., from the free text) as part of a pipeline. The policy decision model was based on prior research, an amalgamation of over 80 user/behavior modeling studies [73].

4.2 AskAPatient

The **AskAPatient** dataset consists of patient demographic information, free-text reviews describing patients' experience with medication, drug duration, and a rating of the medication [60]. In the upstream ML modeling data, there are 20,000 free-text responses. Accounting for missing data in the policy variables, there are approximately 18,000 entries in our dataset. The policy model involved using drug, experience text and ratings, and demographic information as feature variables and usage duration as the downstream prediction variable. This is aligned with prior work that has noted the role of factors such as adverse experiences and online cyberchondria affecting user usage and adherence behaviors [8, 97]. Drug names in the dataset are free-text and, therefore, required standardization. We first converted the brand name drugs to generic names [18]. Then, we identified the **Medical Subject Headings** (**MeSH**) numbers [61] for each drug and categorized them according to the top level (e.g., *D02-Organic Chemicals* for Tylenol). Duration is also a free-text field in the data, (2 weeks, 3 months, etc.). Here we extracted the relevant information and converted all duration values to days for a consistent, numeric scale for our policy model.

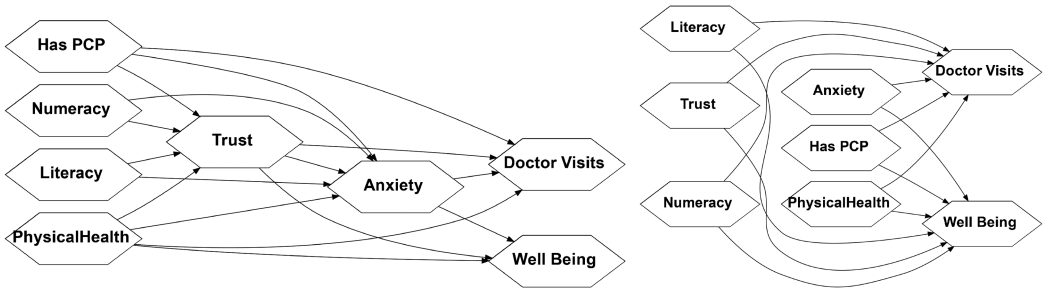
4.3 ML Models

For our upstream datasets, we trained four ML models to predict the output variables. We used two baseline models: BERT [37] and RoBERTa [63]. For each of these models, we fine-tuned the baseline model for the predictive task. More specifically, we used `BERT-base-uncased` and `RoBERTa-base` models from the Huggingface transformer library [98]. BERT and RoBERTa were fine-tuned for five epochs using a batch size of 32, learning rate of $1e^{-5}$, and weight decay of 0.01. Within each of the five folds, the best-performing epoch on the validation set was used for testing. We also ran a debiasing procedure after fine-tuning [51, 52]. For FAIR-Frame interactions, we set $\alpha = 1$. For the fairness regressions, to account for any potential heteroskedasticity concerns, we plotted the residuals and did not observe any megaphone patterns. We also used a robust sandwich estimator and found the effect sizes and statistical significances to be consistent.

The input was the free text provided by the individual users. For Psychometrics, this was the response to the prompt they were given in the mobile survey. For AskAPatient, this was their free-text review of the drug. The variable being predicted for the two tasks was the psychometric survey score and the numeric drug rating, respectively (i.e., a continuous prediction task). For each of these models, in our policy prediction task, we used the output prediction for the variable in place of the gold-standard value. This was done to control for all downstream model factors except the inclusion of the imperfect/erroneous ML prediction scores.

4.4 Downstream Policy Decisions

The motivation for the two downstream model configurations, based on prior literature, was provided in the two dataset overview subsections (Section 4.1 and Section 4.2). For our downstream data, we used the following procedure. First, each policy prediction task includes one feature



(a) Structured Policy Model (b) Linear Policy Model

Fig. 4. Policy models for the downstream prediction tasks. In Figure 4(b), the six input features are staggered for a compact representation.

estimated from an ML model. Each input feature is estimated by one of four possible ML models. Fairness is calculated against one of the available demographic characteristics of interest. For the psychometric data, there are five demographic characteristics: Age, Race, Gender, Income, and Education. For AskAPatient, there are two: Age and Gender.

These estimated input features are then provided in conjunction with additional feature variables of interest to our downstream policy model to predict the decision of interest. In the case of the Psychometric data, we rely on prior work that shows a structural relationship between our feature variables to build a structural equation model [73]. For both tasks, we also employ a "flat" regression model to predict the downstream policy decision (Figure 4). For psychometrics, our policy downstream variables were number of doctor visits in the previous two years and self-reported well being [73]. For AskAPatient, the downstream variable was the duration of drug use.

With this setup, we were able to generate over 5.41 million datapoints. Each datapoint is a unique combination of demographic variable of interest, upstream feature estimated by the ML model, ML model used for estimation, and policy model configuration. For each datapoint we have a gold-standard downstream fairness calculation based on the prediction results from the trained downstream policy model. In the following two sections, we describe how we used these two datasets to evaluate FAIR-Frame and the comparison fairness metrics, and address our two research questions, as follows:

— *Explaining Upstream Fairness with Models* - In Section 5, we illustrate and test the upstream representational fairness of FAIR-Frame versus existing baseline and benchmark metrics. This evaluation relates to our first research question that asks **how different existing fairness metrics are in their measurement of bias in ML representations, and whether fairness models can better measure across multiple protected attributes.**

— *Predicting Downstream Fairness with Models* - In Section 6, we see how well upstream fairness measures align with downstream allocational harm measures from policy/decision models. This evaluation relates to our second research question that asks **how accurately fairness measurement of ML representation can predict bias in downstream tasks/interventions.**

5 EVALUATION: EXPLAINING UPSTREAM FAIRNESS WITH MODELS

In this section, we demonstrate the value of FAIR-Frame for parsimonious explanation of representational fairness. Our upstream evaluation is segmented into two parts. In the first, we statistically compare existing fairness metrics to each other and to FAIR-Frame through an ANOVA

Table 6. ANOVA Results Showing Significant Difference between the Metrics Used to Measure Fairness

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Metric	6	0.39	0.06	20.78	0.0000
UpstreamDV	3	0.15	0.05	15.92	0.0000
Demographic	4	0.43	0.11	34.85	0.0000
Residuals	2,226	6.89	0.00		

(**RQ1a**). In the second part, we illustratively and statistically assess metric-based measures against model-based measurement (i.e., FAIR-Frame) to see how both types hold up for multiple protected attributes (**RQ1b**). In the remainder of the section, we first describe our experimental setup, then compare the results of estimating fairness using the FAIR-Frame model with fairness calculated using existing metrics in the literature (Section 2.2).

5.1 Experimental Setup

First, we trained two ML models with two options for fairness correction for a total of four models. We fine-tuned the BERT and RoBERTa PLMs for each of our testbed tasks, using a five-fold cross-validation setup to generate test predictions for each example in our dataset. For both BERT and RoBERTa, we also employed a post fine-tuning fairness calibration [52]. In total, we trained four PLMs: BERT with no debiasing (**BERT-PT**), BERT with debiasing (**BERT-PTD**), RoBERTa with no debiasing (**RoBERTa-PT**) and RoBERTa with debiasing (**RoBERTa-PTD**). Each model generated a test-set prediction for each example in our testbed. For each of these predictions, we calculated a fairness score using the metrics in our benchmark comparison (see Table 2), as well as with FAIR-Frame.

Our benchmark metrics include both group-level and calibrated metrics. When necessary, we transformed the metric calculation so that all metrics had real-valued outputs. This way we can compare magnitude and direction, in contrast to some metrics (e.g., Gini Coefficient) where the output is always non-negative and directionality of fairness cannot be calculated. For **Demographic Parity**, we do not take the absolute value, but instead, look at the real-value output, so that we can distinguish between disparities that favor the privileged group or the protected group. For the Gini Coefficient (**Gini**), we look at the difference between the Gini Coefficient for all data and the Gini Coefficient for only the protected class. For the calibrated metrics, we make similar modifications. We use the log value of **Adjusted Disparate Impact** (**ADI**). For **Fairness Violation** (**FV**), we use the standard calculation. For Equal Opportunity (**TPRD**), we do not take the absolute value, but instead, use the real-valued output. For the **False Positive Rate Difference** (**FPRD**), we, again, do not take the absolute value, but instead, use the real-valued output. For **Jensen–Shannon Divergence** (**JSD**), we take the difference between the JSD between the true and predicted labels for the privileged class and the JSD between the true and predicted labels for the protected class.

5.2 Results

Table 6 shows the results of an **analysis of variance** (**ANOVA**) to examine how fairness scores vary across metric, PLM, and demographic. We ran an ANOVA to determine if the metric scores were significantly different. Table 6 shows that the metrics are indeed significantly different, suggesting that the choice of fairness measure can impact our view of how fair a given ML model might be. Figure 5 shows the results of a Tukey's test between each pair of metrics. We observe that group consistent metrics devoid of class imbalance/base rate information (i.e., Gini) are significantly different from calibrated metrics, such as JSD and ADI. Moreover, FAIR-Frame is also

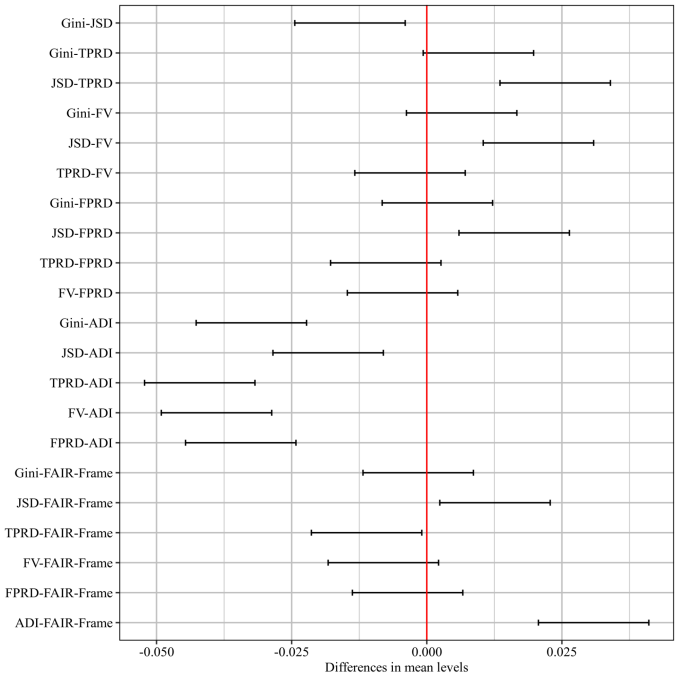


Fig. 5. Mean differences between pairs of fairness metrics as calculated by Tukey's test.

significantly different from several commonly used fairness metrics in the literature, including JSD, TPRD, and ADI. As per **RQ1a**, this result indicates that fairness metrics are not consistent with each other. While each certainly has its own benefits, taken as a whole as part of a data science decision environment, the variability in metrics can drive uncertainty.

We next look at the values themselves to see how the differences manifest in individual instances. Figure 6 plots the various fairness scores across demographic and PLM for the Psychometric data. In the figure, 20 panels are depicted in total. The four rows depict the BERT/RoBERTa PT/PTD models, respectively, whereas the five columns signify the different protected attributes (in this case, demographic variables). For each cell in Figure 6, there is high variance across upstream fairness metrics. Within an organizational context, it is difficult to understand model fairness overall with variety in magnitude and sign across metrics. For example, assessing the fairness of RoBERTa for Age on the Literacy task (Row 3, Column 1, green points in Figure 6) is difficult. Several metrics have negative coefficients, while others have positive coefficients. These inconsistencies again reinforce the need for a consistent representation of fairness that can be deployed in complex decision-making environments.

Next, we closely examine FAIR-Frame outputs. Figures 7 and 8 show the values for each demographic across the Psychometric data tasks. The y-axes show the fairness model coefficient values for how much of the ML model error might be attributable to that respective protected attribute (i.e., the demographic dimensions) presented on the x-axis (along with the fairness model intercepts). Figure 7 depicts the model without any protected attribute interaction effects (i.e., just intercept plus the five demographics), whereas Figure 8 also includes the ten two-way interaction variables for the five demographics.

Taken together, FAIR-Frame provides a holistic, parsimonious representation of fairness across the available demographic characteristics. We can also identify trends in the data, making

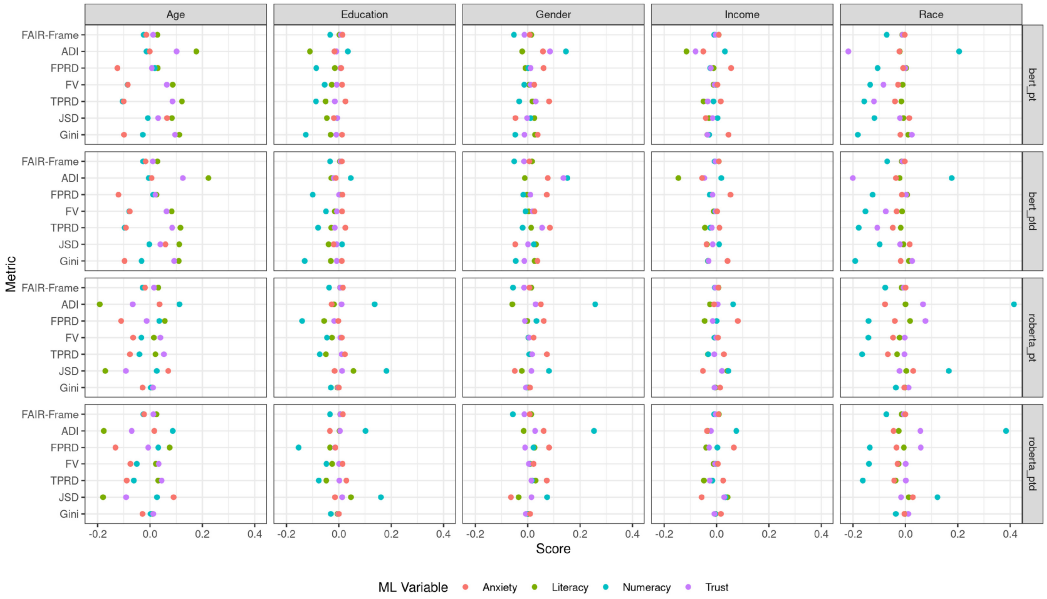


Fig. 6. Error metrics plotted together. FAIR-Frame is consistent across models and does not lead to large swings based on demographics.

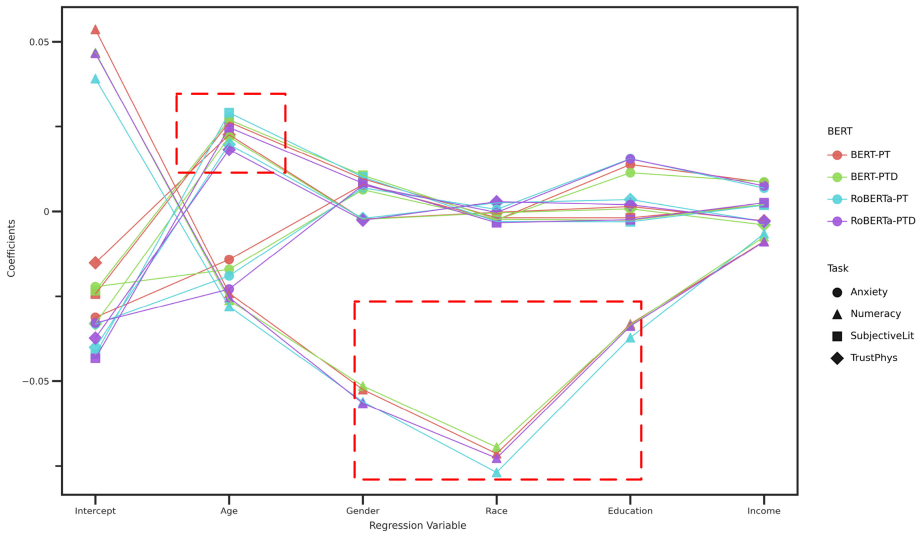


Fig. 7. FAIR-Frame $\hat{\Delta}_R$ with no interactions.

interpretability another value proposition of inference-based modeling approaches such as FAIR-Frame. For example, Figure 7 shows that for the Gender demographic, all four PLMs have a negative coefficient for Numeracy scores for Gender. See the large dashed rectangular region near the bottom of the figure. This means that the estimated error, $\hat{\Delta}_R$ (recall $\Delta_R = z - \hat{z}$), is a larger negative value when the Gender demographic is 1 (i.e., women), indicating that, for women, these models are *overpredicting* numeracy scores (Section 3.2). The y-axis shows the effect size—the overprediction error for women is nearly 5%. This is potentially problematic, as overpredicting

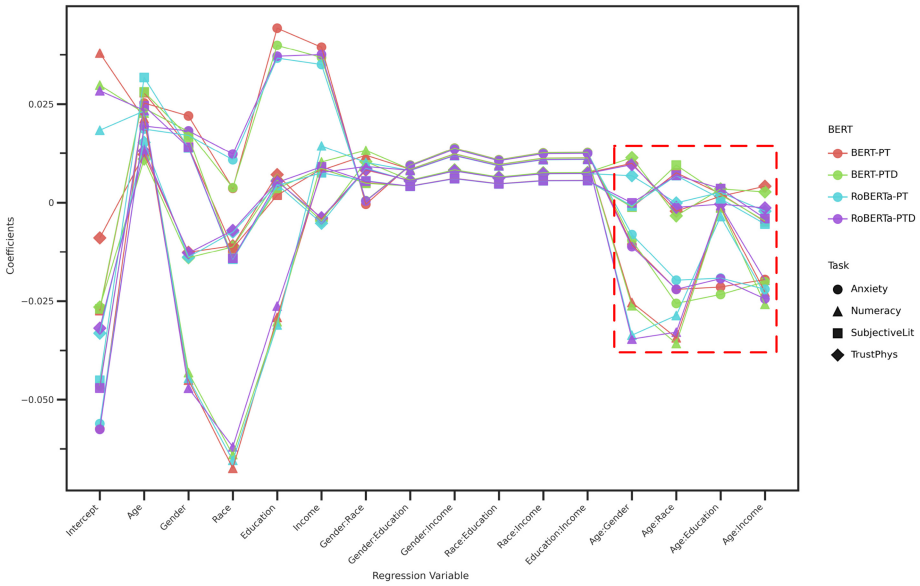


Fig. 8. FAIR-Frame $\hat{\Delta}_R$ with interactions.

numeracy means that women with lower numeracy scores who are in need of an intervention, such as training, may be missed because the PLM model is over-estimating their numeracy. Within that same region of the figure, for protected race attributes, it is even higher, accounting for over-prediction error of 7.5%. Similarly, education attributes also exhibit such an effect on the numeracy task.

On the other hand, the models consistently underpredict subjective literacy for older people. See the smaller dashed rectangular region in the top left of Figure 7. The positive coefficients for Age-Literacy indicate that the estimated error $\hat{\Delta}_R$ is a larger positive value for older people. This indicates that the model predictions are consistently lower than the true values, and the model is underpredicting literacy scores for this group by about 2.5%. This, too, may have negative effects downstream. If the PLM underpredicts numeracy, then unnecessary resources may be spent on these individuals to increase literacy when it is not necessary. These resources could have been better spent on other groups that may need help. In a similar way, looking at Figure 8, we see that models underpredict anxiety for users in two-way protected classes where one of the classes is Age (e.g., older women, older low-income individuals). See the dashed rectangular region in the right side of Figure 8. The figures show how FAIR-Frame is able to parsimoniously show the effect of different protected attributes on ML model error.

We also consider the case of multivariate demographic information. This is often referred to as *intersectional fairness* [92]. Here we consider two-way demographic information (e.g., young males and older females). We follow prior work [57] and look at the differences between two-way protected groups (e.g., non-white females) and their demographic negations (white males). This way we do not simply aggregate the one-way metrics, but instead, look at the gaps between two specific groups.

Figure 9 shows the difference between ADI and FAIR-Frame across models and demographics for the four upstream variables from the Psychometric dataset. The heights of the bars indicate the magnitude of observed bias, whereas the placement above/below the x-axis denotes the direction of the ML model bias effect. As illustrated in the figure, FAIR-Frame values are more stable than

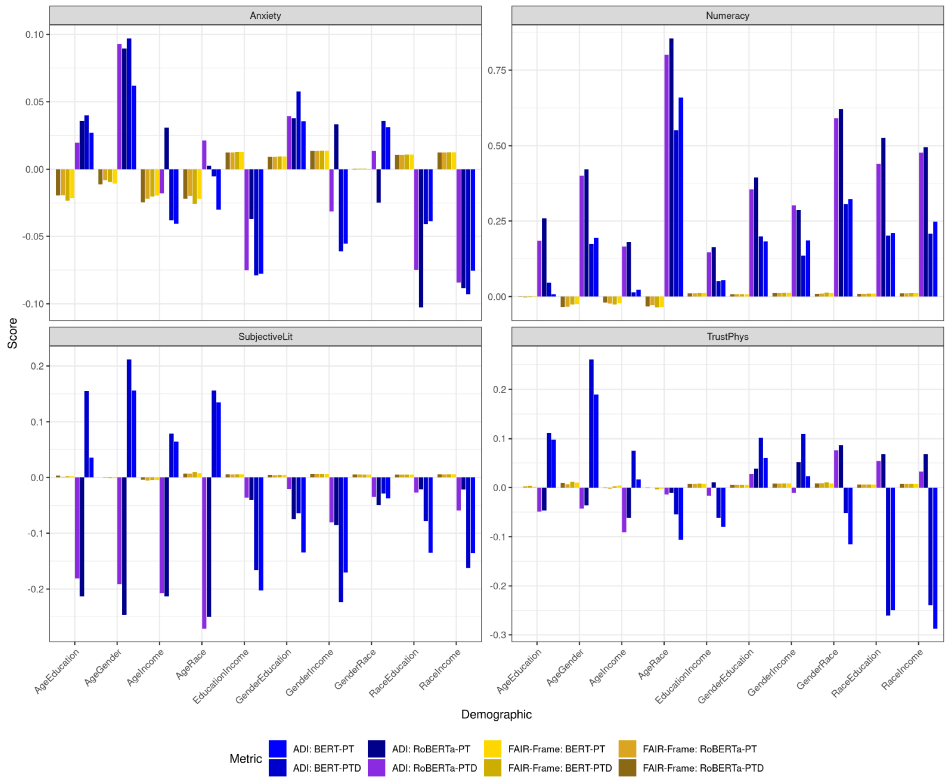


Fig. 9. Comparing ADI and FAIR-Frame in the case of intersectional demographics.

ADI. First, FAIR-Frame values are typically much closer to zero and offer a more realistic quantification than ADI. For example, ADI scores for the Age-Race pair on the Numeracy task are between 0.5 and 0.75, which are large numbers considering our fair baseline is 0. Second, FAIR-Frame values exhibit consistency across the ML models used to generate the predictions, where ADI values fluctuate model-to-model (**RQ1b**). For example, for two way demographics that include Age on the Subjective Literacy task, BERT and RoBERTa predictions lead to opposing signs for ADI scores. See bars in the left region of the bottom left panel in Figure 9. The prevalence of large magnitude, high variance, and unstable direction effects observed with ADI is consistent with our theoretical analysis on unpredictable bias (Section 3.3.3) and underscores a bevy of practical usage issues and concerns for such model-free metrics. As mentioned above, these variations make already complex decisions regarding model selection even more complex. If it is unclear how fair certain models are for representational tasks, then that uncertainty could propagate or cascade down to the allocational task as well [55]. Based on our analyses of representational fairness, we have shown that there are significant differences in fairness metric outputs (**RQ1a**) and that FAIR-Frame is able to model fairness more consistently than existing metrics, especially with regard to intersectional cases (**RQ1b**). Next, we consider the predictive power of these fairness measurements on downstream allocational fairness.

6 EVALUATION: PREDICTING DOWNSTREAM FAIRNESS WITH MODELS

In this section, we analyze the predictive power of each representational fairness metric with regard to the fairness of downstream policy predictions. For each of our datasets, there are one

Table 7. Psychometric NLP Task Results Table

	Method	Alignment						Magnitude Difference					
		LIT	NUM	TR	ANX	Avg	Lift	LIT	NUM	TR	ANX	Avg	Lift
Doc Visits													
Linear													
	FAIR-Frame	**0.60**	**0.60**	**0.60**	0.75	**0.64**	0.68	0.068	**0.062**	0.065	0.063	**0.064**	−0.45
	ADI	0.50	0.50	0.35	0.25	0.40	0.05	0.083	0.175	0.115	0.095	0.117	–
	FPRD	0.25	0.35	0.35	0.70	0.41	0.08	0.086	0.112	0.080	**0.047**	0.081	−0.31
	FV	0.20	0.50	0.25	**0.80**	0.44	0.16	0.079	0.080	0.082	0.052	0.073	−0.38
	TPRD	0.20	0.50	0.25	**0.80**	0.44	0.16	0.095	0.090	0.097	0.057	0.085	−0.27
	JSD	0.50	0.55	0.55	0.20	0.45	0.18	**0.065**	0.104	**0.058**	0.091	0.080	−0.32
	Gini	0.10	0.50	0.20	0.70	0.38	–	0.088	0.088	0.085	0.055	0.079	−0.32
Structured													
	FAIR-Frame	**0.60**	**0.60**	**0.60**	0.75	**0.64**	0.68	0.071	**0.065**	0.061	0.059	**0.064**	−0.45
	ADI	0.50	0.50	0.35	0.25	0.40	0.05	0.086	0.177	0.112	0.091	0.116	–
	FPRD	0.25	0.35	0.35	0.70	0.41	0.08	0.089	0.117	0.076	0.049	0.083	−0.28
	FV	0.20	0.50	0.25	**0.80**	0.44	0.16	0.083	0.084	0.080	**0.048**	0.074	−0.36
	TPRD	0.20	0.50	0.25	**0.80**	0.44	0.16	0.101	0.095	0.094	0.054	0.086	−0.26
	JSD	0.50	0.55	0.55	0.20	0.45	0.18	**0.065**	0.106	**0.055**	0.087	0.078	−0.33
	Gini	0.10	0.50	0.20	0.70	0.38	–	0.091	0.092	0.081	0.050	0.079	−0.32
Well Being													
Linear													
	FAIR-Frame	**0.80**	0.40	0.40	0.95	**0.64**	0.68	0.145	0.153	0.144	0.135	**0.144**	-0.18
	ADI	0.30	0.70	0.55	0.45	0.50	0.32	0.176	0.199	0.174	0.152	0.175	–
	FPRD	0.15	0.45	0.35	0.90	0.46	0.21	0.167	0.179	0.152	**0.098**	0.149	−0.15
	FV	0.40	0.50	0.45	**1.00**	0.59	0.55	0.159	0.152	0.155	0.117	0.146	−0.17
	TPRD	0.40	0.50	0.45	**1.00**	0.59	0.55	0.172	0.164	0.162	0.112	0.152	−0.13
	JSD	0.50	**0.75**	**0.65**	0.00	0.48	0.26	**0.142**	**0.130**	**0.130**	0.186	0.147	−0.16
	Gini	0.30	0.30	0.00	0.90	0.38	–	0.166	0.168	0.161	0.120	0.154	−0.12
Structured													
	FAIR-Frame	**0.80**	0.40	0.35	0.90	**0.61**	0.61	0.154	0.153	0.151	0.122	**0.145**	−0.18
	ADI	0.30	0.70	0.50	0.40	0.48	0.26	0.184	0.200	0.185	0.139	0.177	–
	FPRD	0.15	0.45	0.30	0.85	0.44	0.16	0.177	0.179	0.160	**0.092**	0.152	−0.14
	FV	0.40	0.50	0.40	**0.95**	0.56	0.47	0.169	0.152	0.161	0.104	0.146	−0.18
	TPRD	0.40	0.50	0.40	**0.95**	0.56	0.47	0.180	0.164	0.170	0.098	0.153	−0.14
	JSD	0.50	**0.75**	**0.60**	0.05	0.48	0.26	**0.148**	**0.141**	**0.137**	0.172	0.149	−0.16
	Gini	0.30	0.30	0.05	0.85	0.38	–	0.174	0.170	0.167	0.108	0.155	−0.12

Alignment is calculated as the percentage of instances where representational and allocational fairness are aligned.

or more downstream predictions that can be made using the fine-tuned PLM *outputs* as *inputs*. For example, in the case of the Psychometric dataset, each of the four estimated psychometric dimensions are inputs to a downstream model to predict subjective well-being (**Well Being**) or the number of times a patient has visited the doctor (**Doc Visits**) [9, 73]. For AskAPatient, the estimated rating is an input to predict the duration of drug use (**Duration**). What's more, prior work in the literature [3, 73] shows a structural relationship across inputs for predicting **Doc Visits** and **Well Being**, which we leverage in our downstream model formulation (Figure 4(a)).

We set up our downstream evaluation as follows. To start, all input features for the downstream prediction task were fixed. For each upstream variable predicted by a PLM (e.g., Anxiety), we replaced the true feature (i.e., the scored psychometric result) with the estimate from the PLM. With this modified feature set, we trained our downstream policy model to generate predictions for the downstream task (e.g., Well Being). Gold standard fairness for the downstream task was ADI. We binarized the downstream tasks at the median true value for calculating fairness. To calculate alignment, we compared the sign of each upstream fairness metric with the sign of the downstream ADI. We also calculated the magnitude difference between the upstream and downstream metrics.

Tables 7 and 8 show how FAIR-Frame compares to other metrics when used to predict downstream fairness. The four columns labeled LIT, NUM, TR, and ANX denote the alignment or

Table 8. AskAPatient Rating Task Results Table

PolicyModel	Method	Alignment	Lift	MagnitudeDifference	Lift
Linear	FAIR-Frame	**0.88**	6.33	0.044	−0.48
	ADI	0.12	0.00	0.084	–
	FPRD	0.50	3.17	0.048	−0.43
	FV	0.12	0.00	0.055	−0.35
	TPRD	0.12	–	0.071	−0.15
	JSD	0.50	3.17	0.050	−0.40
	Gini	0.50	3.17	**0.039**	−0.54

magnitude difference on the literacy, numeracy, trust, and anxiety tasks, respectively. The column Avg is the average across the four tasks, whereas lift is the percentage improvement over the Gini baseline. For the Psychometrics data (Table 7), we see that on average across the upstream variables, FAIR-Frame provides the best alignment in terms of predicting the sign of the downstream fairness metric (Avg alignment of 61–64%) and also has the lowest average difference in value. On three of the four downstream settings, the alignment is about 20 points higher than the best benchmark fairness metrics—JSD, TPRD, and FV—none of which surpass 50% in alignment. On the structured well-being setting, the performance deltas are a bit less pronounced, but nevertheless, FAIR-Frame attains gains of 5 points in terms of alignment over the best comparison metrics. In regards to percentage improvement, compared to the Gini Coefficient measurement for alignment, FAIR-Frame has an improvement of over 60%. For magnitude, FAIR-Frame improves over ADI by as much as 45%. These measures are important since the downstream model thresholds are the basis for various policy outcomes with economic and societal implications [55].

For anxiety (ANX), we find that FAIR-Frame does slightly worse than other models, in particular FV. Upon inspection, we found that this was due to cases where FAIR-Frame was misaligned with a small coefficient, while FV was aligned with a large coefficient. This is consistent with our theoretical results (Section 3.3.3), which state that the unobserved variable bias can lead to unpredictable biases, which in practice means that there may be certain cases (such as ANX here), where the biased estimate is aligned while FAIR-Frame is not. Because the existing methods can be either positively or negatively biased estimates, there are instances where by chance, select comparison methods attain smaller magnitude differences (or better alignment) as compared to FAIR-Frame. This results in certain comparison methods outperforming FAIR-Frame in one-off cases. Case in point, FPRD, FV, and TPRD have large upstream scores for age on the anxiety dependent variable (see the first cell in Figure 6 appearing earlier) that happen to coincide with the downstream allocational harm magnitude/direction. However, because these magnitudes/alignments are random, they appear sporadically across comparison methods in Table 6, and on average, FAIR-Frame yields markedly better results compared to all benchmark and baseline methods. The latter point is evidenced by the overall (i.e., the Avg column) in Table 6, where we find that FAIR-Frame is more aligned with downstream fairness than our benchmark metrics.

For AskAPatient, as shown in Table 8, FAIR-Frame has the best alignment and second-best difference in magnitude. Once again, none of the comparison metrics surpass 50% alignment, suggesting odds no better than a coin flip for inferring downstream allocational harm based on their representational harm directionality. In contrast, FAIR-Frame attains 88% alignment, suggesting high alignment and greater potential for predicting how the representational harm profile of a given ML model may relate to its allocational implications. With respect to percentage improvement over the ADI, FV, and TPRD metrics yielding the lowest alignment, FAIR-Frame is more than 6 times higher in alignment.

Figure 10 summarizes the average alignment and magnitude difference across downstream variables for each fairness metric on the Psychometric dataset. As noted, the gap between FAIR-Frame

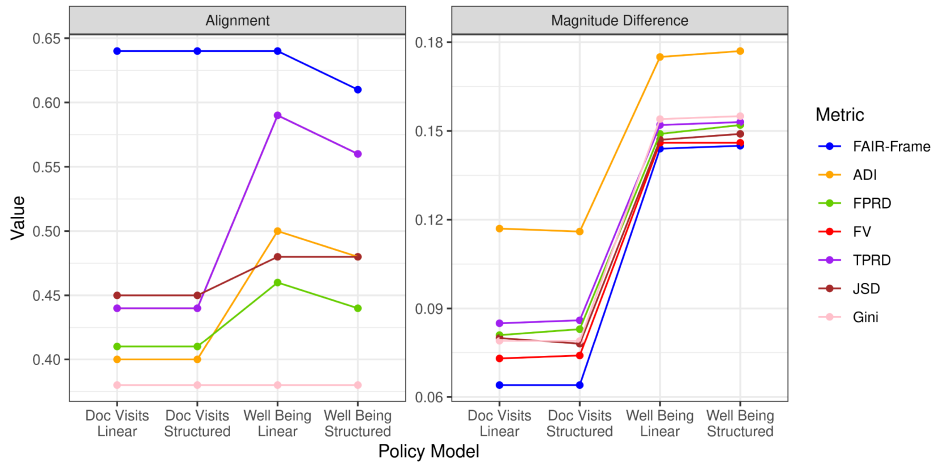


Fig. 10. Overall performance of each metric across ML model estimates.

and the other metrics is large, especially for the **Doc Visits** downstream variable. As mentioned above, this predictive element of representational fairness is important when considering the broader environment in which such metrics will be deployed and analyzed by humans. If a data scientist uses the (typically) more readily accessible representational fairness metric as a proxy for allocational fairness when deciding which algorithm to use, then misalignment between representational and allocational fairness can lead to unexpected downstream harms that might be problematic as ML modeling moves towards responsible and trustworthy AI frameworks [59]. These results further underscore the importance of model-based fairness inference and call into question the efficacy of some of the fairness metrics proposed in the literature. Although such metrics seem reasonable, intuitive, and easy to use, our findings at the interplay of upstream and downstream in complex real-world scenarios suggest that perhaps certain existing metrics lack practical robustness.

The results presented in Tables 7 and 8, and summarized in Figure 10, are aggregated across underlying models and demographics (n=20 for LIT, NUM, TR, and ANX; n=80 for Avg). In order to explore the performance across demographics, and by model types, we aggregated the results within these respective dimensions. Figure 11 shows the results by demographics for FAIR-Frame and the six comparison metrics across the two policy outcome variables. Each x-axis group was aggregated across the four tasks and four models (i.e., n=16). Based on the alignment scores, FAIR-Frame was as good or better than all comparison metrics on 9 of the 10 demographics. The one exception was on Gender for the well-being policy outcome, where it underperformed the other metrics. Similarly, Figure 12 shows the results by model type. Each x-axis group was aggregated across the four tasks and five demographics (i.e., n=20). Once again, FAIR-Frame attained the best alignment percentages on 7 of the 8 cases, being marginally outperformed by FV, JSD, and TPRD on the RoBERTa-PT model. Nevertheless, overall, the results lend further credence to the notion that model-based fairness might be more effective for alignment as opposed to metric-based measures.

Collectively, the prediction results further support the efficacy and utility of model-based measurement of ML bias through frameworks such as FAIR-Frame. Accurate prediction of upstream-downstream fairness alignment and error is important. Misalignment between upstream and downstream metrics can have implications in our testbed context regarding delivery of user-centric personalized interventions. If an upstream unfairness (against e.g., older individuals) is measured, and the expectation is that this value persists downstream, interventions may be made

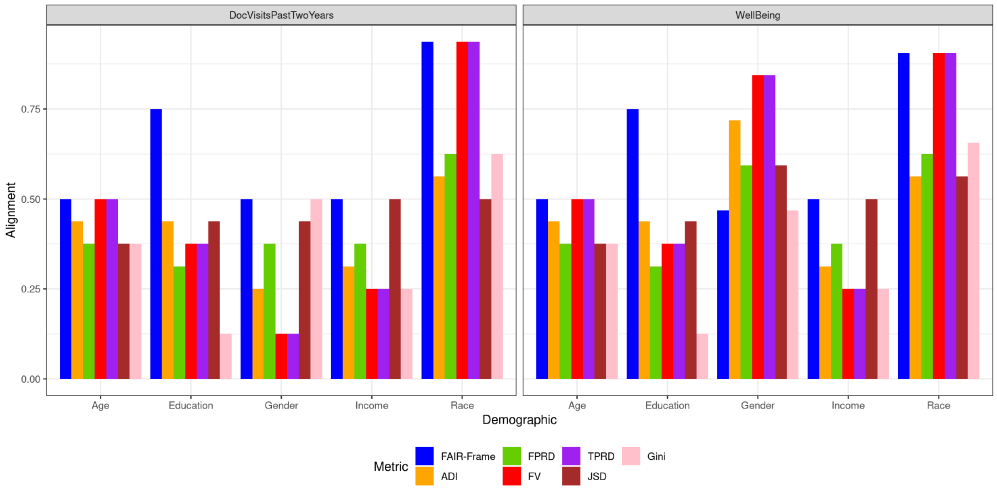


Fig. 11. Alignment of each fairness metric, grouped by demographic.

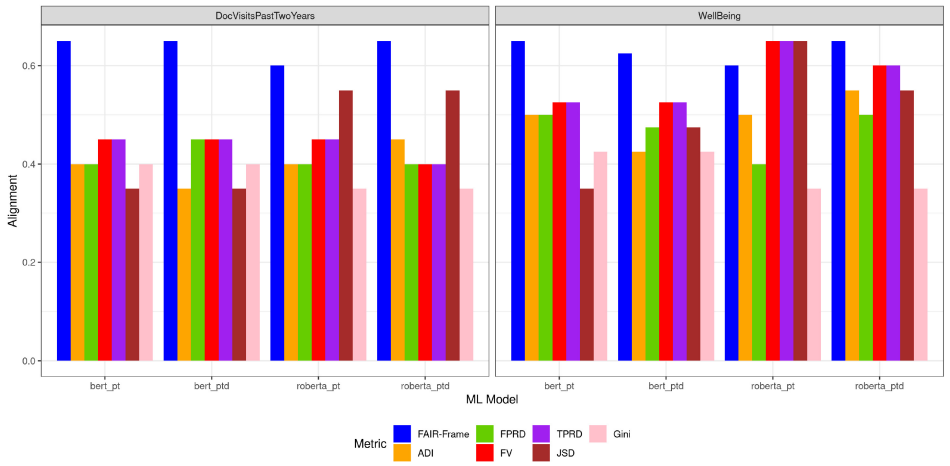


Fig. 12. Alignment of each fairness metric, grouped by NLP model.

to encourage more older individuals to visit the doctor. However, if the signs are not aligned, then the unfairness downstream is actually against younger individuals, and the resources spent on encouraging this behavior in the elderly should have been instead targeting younger people. Similarly, just as sign/directionality alignment matters, the error in effect sizes can also have unintended consequences for well-intentioned decisions and policies—it can over/estimate the extend of disparities due to ML error/bias. Hence, although it is critical to understand representational biases, predicting allocational fairness can help mitigate unfairness in implementations of systems that make policy decisions and route resources to different groups [19, 55].

7 ANALYSIS: DEBIASING WITH FAIR-FRAME

As a final analysis, we look to see if FAIR-Frame's predictive power can be leveraged for debiasing as well. As noted, debiasing is not the goal for fairness measurement. However, one possible

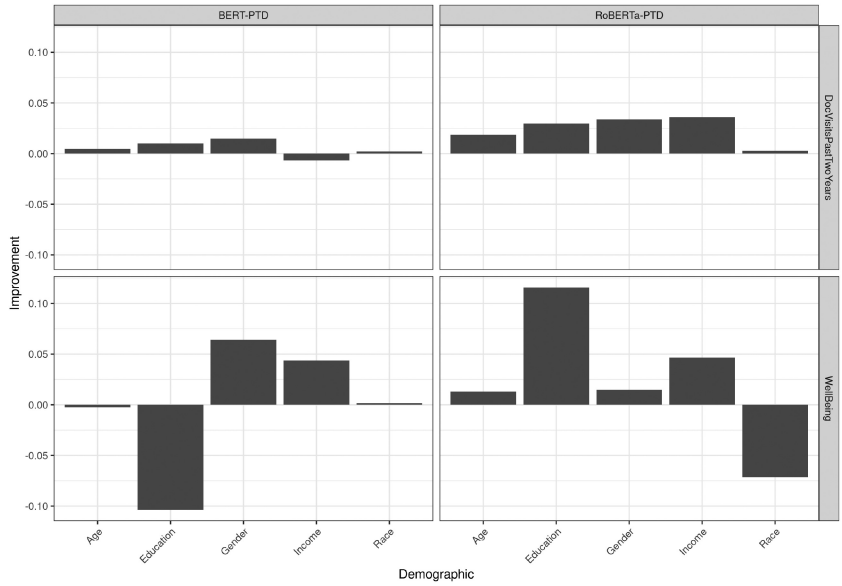


Fig. 13. Debiasing using FAIR-Frame coefficients can improve the fairness of downstream policy models.

future direction for model-based fairness could be to embed such models as part of an end-to-end fairness by design strategy [5]. Here, we experiment with a simple debiasing setup involving the use of the fairness measures derived with FAIR-Frame as an ex-post correction mechanism on the $\hat{z}$ predictions from the upstream ML models. To illustrate this, we focus on Health Numeracy from the Psychometric testbed as the downstream policy variable. Given the outputs of FAIR-Frame and the upstream model predictions (e.g., BERT-PTD) for Health Numeracy, we adjust each prediction based on whether the record is associated with the protected class in each of the demographic groups:

$$\hat{z}_{\text{FF}} = \hat{z} - \sum_{d \in D} \mathbb{I}[d = 1] * FF_d, \tag{29}$$

where d signifies a protected attribute such as age, income, and so on, and FF_d is the coefficient for that particular protected attribute from FAIR-Frame. That is, FF_d is the β_d estimated for X_d in Equation (9). Recall that a positive value for FAIR-Frame coefficients means that the model is overpredicting for that protected group. Therefore, we subtract the sum of the coefficients to account for the correct direction of the adjustment.

We intentionally focused on the BERT-PTD and RoBERTa-PTD models since both of these were already debiased after pretraining (that is, the model weights were adjusted before fine-tuning). Further debiasing a PLM after fine-tuning is considered an acceptable strategy to avoid upstream representational harm, and possibly, for reducing downstream allocational harm [57]. The results of our illustrative debiasing analysis exercise appear in Figure 13. The bars indicate the amount of reduction in downstream allocational harm attributable to our simple ex-post correction on the upstream NLP models' predictions. More specifically, the plot shows positive values when FAIR-Frame debiasing improves fairness (i.e., a value closer to 0 after debiasing) and negative values when FAIR-Frame does not improve fairness (i.e., a value further away from 0). We see that across the 20 demographic-model tuple combinations, the $\hat{z}_{\text{FF}}$ correction reduces bias in 17 out of 20 situations (i.e., 85% of the combinations). This is markedly higher than the results attained in prior fine-tuned model debiasing scenarios involving interactional effects across multiple protected

attributes [57, 92]. Admittedly, the results do worsen bias in a few cases, most notably, on the Well Being policy variable for Education when using BERT-PTD, and Race, when using RoBERTa-PTD. However, we believe such a debiasing strategy, with further development such as considering adjustment effects for different protected attributes could be promising. Future work investigating FAIR-Frame as a possible debiasing mechanism, or broadly considering model-based fairness as part of an end-to-end fair learning strategy or a post-hoc adjustment procedure may represent worthwhile avenues.

8 CONCLUSION

With the proliferation of ML models for process automation and augmentation in an array of contexts, the ability to measure ML fairness has come front and center. However, fairness measurement has encountered a few challenges. First, the impetus has been to measure using simple, intuitive univariate metrics that are well-intentioned, but fail to scale to multi-variate settings rife with many protected attributes and interaction effects. Second, most prior work fails to evaluate the alignment of upstream metrics with their implications for downstream allocational harm within pipelines of ML models. Third, fairness measurement research assumes a sandbox environment that is disentangled from the ML modeling process—failing to consider the grid/random search aspect of model selection, or the need to apply fairness measures to multiple models. Collectively, these limitations produce fairness metrics that are less effective and a difficult smorgasbord for decision-making.

In order to tackle these challenges, we propose FAIR-Frame: a framework that considers the allocational, interactional, and representational aspects of fairness. FAIR-Frame encompasses an integrated model that can measure fairness in upstream models (i.e., representational harm) in a manner that is reasonably predictive of the fairness alignment in downstream allocational harm attributable to policy/decision outcomes involving the ML model. Experiments on two testbeds encompassing over 5.41 million combined upstream and downstream observations spanning 27 thousand users and over 50 thousand documents highlight the effectiveness of FAIR-Frame relative to various existing fairness metrics. In particular, FAIR-Frame parsimoniously measures multivariate and interactional bias and has less variance in its measurement of bias across protected attributes in upstream models—producing measures that are significantly different from certain group consistent and calibrated metrics. Moreover, FAIR-Frame's upstream measures have a 10 to 40 point lift in alignment with downstream allocational outcomes, while producing comparable or lower mean errors than benchmark metrics. Consistent with prior work, our primary contributions include the framework [9, 89], empirical insights attained [11, 95], and our extensive evaluation on two rich testbeds comprising upstream and downstream contexts.

The results of our work have important implications for research related to the design, development, and implementation of fair ML strategies in organizational responsible AI environments [2]. First, our results show the importance of moving beyond univariate metric perspectives of fairness. ML modeling and policy/decision spaces are multi-dimensional, multi-objective, and interactional. For instance, text/NLP models, new recommender systems, novel search algorithms, user modeling strategies, and legal judgment predictions, to name a few, all embody these characteristics [96, 100, 106]. In such spaces, there is a need to parsimoniously consider interactions and align fairness objectives/metrics with the practicalities and constraints of the ML modeling process [71, 80]. Second, our findings underscore the importance of aligning upstream representational fairness with downstream allocational harm [17, 19]. Namely, we show the efficacy of a fairness modeling design that considers the explanation and prediction perspectives in an integrated way [50]. Our results also have implications for practice, such as embedding fairness by design through flexible,

adaptive models that can holistically measure fairness in grid/random search spaces [5]. Moreover, although our evaluation focused on downstream contexts where a single upstream model was deployed at a time, future work may consider extensions that shed light on the impact of downstream environments involving multiple models deployed at once (i.e., lengthier fairness pipelines). Additionally, our work focused on two NLP text sequence classification problems—however, the proposed framework is ML model and data/type agnostic since it relies on the predictions and their downstream usage via different types of weighted decision-making scenarios. Moreover, although FAIR-Frame uses highly interpretable models for measuring representational bias [82], future work could explore more complex models with different prediction-explanation tradeoffs. Although future research opportunities abound, we believe our work signifies an important step towards more parsimonious and holistic perspectives of fairness measurement in ML.

REFERENCES

[1] Ahmed Abbasi, Hsinchun Chen, and Arab Salem. 2008. Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. *ACM Transactions on Information Systems* 26, 3 (2008), 1–34.

[2] Ahmed Abbasi, Roger H. L. Chiang, and Jennifer Xu. 2023. Data science for social good. *Journal of the Association for Information Systems* 24, 6 (2023), 1439–1458.

[3] Ahmed Abbasi, David Dobolyi, John P. Lalor, Richard G. Netemeyer, Kendall Smith, and Yi Yang. 2021. Constructing a psychometric testbed for fair natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 3748–3758.

[4] Ahmed Abbasi, Stephen France, Zhu Zhang, and Hsinchun Chen. 2010. Selecting attributes for sentiment classification using feature relation networks. *IEEE Transactions on Knowledge and Data Engineering* 23, 3 (2010), 447–462.

[5] A. Abbasi, J. Li, G. Clifford, and H. Taylor. 2018. Make 'fairness by design' part of machine learning. *Harvard Business Review*, August 1. https://hbr.org/2018/08/makefairness-by-design-part-of-machine-learning

[6] Ahmed Abbasi, Suprateek Sarker, and Roger H. L. Chiang. 2016. Big data research in information systems: Toward an inclusive research agenda. *Journal of the Association for Information Systems* 17, 2 (2016), 3.

[7] Ajay Agrawal, Joshua Gans, and Avi Goldfarb. 2018. *Prediction Machines: The Simple Economics of Artificial Intelligence*. Harvard Business Press.

[8] Faizan Ahmad, Ahmed Abbasi, Brent Kitchens, Donald Adjeroh, and Daniel Zeng. 2022. Deep learning for adverse event detection from web search. *IEEE Transactions on Knowledge and Data Engineering* 34, 6 (2022), 2681–2695.

[9] Faizan Ahmad, Ahmed Abbasi, Jingjing Li, David G. Dobolyi, Richard G. Netemeyer, Gari D. Clifford, and Hsinchun Chen. 2020. A deep learning architecture for psychometric natural language processing. *ACM Transactions on Information Systems* 38, 1 (2020), 1–29.

[10] H. Akaike. 1973. Information theory and an extension of the maximum likelihood principle. In *Proceedings of the 2nd International Symposium on Information Theory, 1973*. Akademiai Kiado.

[11] Jaime Arguello and Bogeum Choi. 2019. The effects of working memory, perceptual speed, and inhibition in aggregated search. *ACM Transactions on Information Systems* 37, 3 (2019), 1–34.

[12] Yazeed Awwad, Richard Fletcher, Daniel Frey, Amit Gandhi, Maryam Najafian, and Mike Teodorescu. 2020. *Exploring Fairness in Machine Learning for International Development*. Technical Report. CITE MIT D-Lab.

[13] Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. 2017. The problem with bias: Allocative versus representational harms in machine learning. In *Proceedings of the 9th Annual Conference of the Special Interest Group for Computing, Information and Society*.

[14] Solon Barocas and Andrew D. Selbst. 2016. Big data's disparate impact. *California Law Review* 104, 3 (2016), 671–732.

[15] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 610–623.

[16] James Bergstra and Yoshua Bengio. 2012. Random search for hyper-parameter optimization. *Journal of Machine Learning Research* 13 (2012), 281–305.

[17] Richard A. Berk, Arun Kumar Kuchibhotla, and Eric Tchetgen Tchetgen. 2022. Fair risk algorithms. *Annual Review of Statistics and Its Application* 10 (2022), 165–187.

[18] Michael L. Bernauer. 2017. Mlbernauer/drugstandards: Python library for standardizing drug names (v0.1). Zenodo. https://doi.org/10.5281/zenodo.571248

[19] S. L. Blodgett, S. Barocas, H. Daumé III, and H. Wallach. 2020. Language (Technology) is power: A critical survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 5454–5476.

[20] Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of african-american english. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. 1119–1130.

[21] Kenneth A. Bollen and Mark D. Noble. 2011. Structural equation models and the quantification of behavior. *Proceedings of the National Academy of Sciences* 108, supplement_3 (2011), 15639–15646.

[22] Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T. Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in Neural Information Processing Systems* 29 (2016).

[23] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, S. Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen A. Creel, Jared Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren E. Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas F. Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, O. Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir P. Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Benjamin Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, J. F. Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Robert Reich, Hongyu Ren, Frieda Rong, Yusuf H. Roohani, Camilo Ruiz, Jack Ryan, Christopher R'e, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishna Parasuram Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei A. Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2021. On the opportunities and risks of foundation models. ArXiv (2021). Retrieved from https://crfm.stanford.edu/assets/report.pdf

[24] Avishek Bose and William Hamilton. 2019. Compositional fairness constraints for graph embeddings. In *Proceedings of the International Conference on Machine Learning*. PMLR, 715–724.

[25] Amanda Bower, Sarah N. Kitchen, Laura Niss, Martin J. Strauss, Alexander Vargas, and Suresh Venkatasubramanian. 2017. Fair pipelines. In *Proceedings of the Workshop on Fairness, Accountability, and Transparency in Machine Learning (FAT/ML)*.

[26] Donald E. Brown, Ahmed Abbasi, and Raymond Y. K. Lau. 2015. Predictive analytics: Predictive modeling at the micro level. *IEEE Intelligent Systems* 30, 3 (2015), 6–8.

[27] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the Conference on Fairness, Accountability and Transparency*. PMLR, 77–91.

[28] Robin Burke. 2017. Multisided fairness for recommendation. 2017 *Workshop on Fairness, Accountability, and Transparency in Machine Learning (FAT/ML'17)*.

[29] Gordon Burtch, Yili Hong, Ravi Bapna, and Vladas Griskevicius. 2018. Stimulating online reviews by combining financial incentives and social norms. *Management Science* 64, 5 (2018), 2065–2082.

[30] Ángel Alexander Cabrera, Will Epperson, Fred Hohman, Minsuk Kahng, Jamie Morgenstern, and Duen Horng Chau. 2019. FairVis: Visual analytics for discovering intersectional bias in machine learning. In *Proceedings of the 2019 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE, 46–56.

[31] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334 (2017), 183–186.

[32] Jacqueline G. Cavazos, P Jonathon Phillips, Carlos D. Castillo, and Alice J. O'Toole. 2020. Accuracy comparison across face recognition algorithms: Where are we on measuring race bias? *IEEE Transactions on Biometrics, Behavior, and Identity Science* 3, 1 (2020), 101–111.

[33] Tessa E. S. Charlesworth, Aylin Caliskan, and Mahzarin R. Banaji. 2022. Historical representations of social groups across 200 years of word embeddings from Google Books. *Proceedings of the National Academy of Sciences* 119, 28 (2022), e2121798119.

[34] Jiawei Chen, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, and Xiangnan He. 2023. Bias and debias in recommender system: A survey and future directions. *ACM Trans. Inf. Syst.* 41, 3, Article 67 (July 2023), 39 pages. https://doi.org/10.1145/3564284

[35] Alexandra Chouldechova and Aaron Roth. 2020. A snapshot of the frontiers of fairness in machine learning. *Communications of the ACM* 63, 5 (2020), 82–89.

[36] Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 120–128.

[37] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. DOI: https://doi.org/10.18653/v1/N19-1423

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

[39] Thomas Elsken, Jan Hendrik Metzen, and Frank Hutter. 2019. Neural architecture search: A survey. *The Journal of Machine Learning Research* 20, 1 (2019), 1997–2017.

[40] Vitalii Emelianov, George Arvanitakis, Nicolas Gast, Krishna P. Gummadi, and Patrick Loiseau. 2019. The price of local fairness in multistage selection. In *Proceedings of the IJCAI-2019-28th International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization, 5836–5842.

[41] Sorelle A. Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P. Hamilton, and Derek Roth. 2019. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 329–338.

[42] Batya Friedman and Helen Nissenbaum. 1996. Bias in computer systems. *ACM Transactions on Information Systems* 14, 3 (1996), 330–347.

[43] Tianjun Fu, Ahmed Abbasi, Daniel Zeng, and Hsinchun Chen. 2012. Sentimental spidering: Leveraging opinion information in focused crawlers. *ACM Transactions on Information Systems* 30, 4 (2012), 1–30.

[44] Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences* 115, 16 (2018), E3635–E3644.

[45] Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. 2021. Intrinsic bias metrics do not correlate with application bias. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 1926–1940.

[46] Yue Guo, Yi Yang, and Ahmed Abbasi. 2022. Auto-debias: Debiasing masked language models with automated biased prompts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1012–1023.

[47] Xiangnan He, Zhaochun Ren, Emine Yilmaz, Marc Najork, and Tat-Seng Chua. 2021. Graph technologies for user modeling and recommendation: Introduction to the special issue - part 1. *ACM Transactions on Information Systems* 40, 2, Article 21 (2021), 5 pages. DOI: https://doi.org/10.1145/3477596

[48] Xiangnan He, Zhaochun Ren, Emine Yilmaz, Marc Najork, and Tat-Seng Chua. 2021. Introduction to the special section on graph technologies for user modeling and recommendation, part 2. *ACM Transactions on Information Systems* 40, 3, Article 42 (2021), 5 pages. DOI: https://doi.org/10.1145/3490180

[49] Xin He, Kaiyong Zhao, and Xiaowen Chu. 2021. AutoML: A survey of the state-of-the-art. *Knowledge-Based Systems* 212 (2021), 106622.

[50] Jake M. Hofman, Duncan J. Watts, Susan Athey, Filiz Garip, Thomas L. Griffiths, Jon Kleinberg, Helen Margetts, Sendhil Mullainathan, Matthew J. Salganik, Simine Vazire, Alessandro Vespignani, and Tal Yarkoni. 2021. Integrating explanation and prediction in computational social science. *Nature* 595, 7866 (2021), 181–188.

[51] Masahiro Kaneko and Danushka Bollegala. 2019. Gender-preserving debiasing for pre-trained word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 1641–1650.

[52] Masahiro Kaneko and Danushka Bollegala. 2021. Debiasing pre-trained contextualised embeddings. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. 1256–1266.

[53] Svetlana Kiritchenko and Saif Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of the 7th Joint Conference on Lexical and Computational Semantics*. Association for Computational Linguistics, New Orleans, Louisiana, 43–53. DOI: https://doi.org/10.18653/v1/S18-2005

[54] Brent Kitchens, David Dobolyi, Jingjing Li, and Ahmed Abbasi. 2018. Advanced customer analytics: Strategic value through integration of relationship-oriented big data. *Journal of Management Information Systems* 35, 2 (2018), 540–574.

[55] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. 2015. Prediction policy problems. *American Economic Review* 105, 5 (2015), 491–495.

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

[57] John Lalor, Yi Yang, Kendall Smith, Nicole Forsgren, and Ahmed Abbasi. 2022. Benchmarking intersectional biases in NLP. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Seattle, United States, 3598–3609. DOI: https://doi.org/10.18653/v1/2022.naacl-main.263

[58] Min Kyung Lee, Anuraag Jain, Hea Jin Cha, Shashank Ojha, and Daniel Kusbit. 2019. Procedural justice in algorithmic fairness: Leveraging transparency and outcome control for fair algorithmic mediation. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–26.

[59] Bo Li, Peng Qi, Bo Liu, Shuai Di, Jingen Liu, Jiquan Pei, Jinfeng Yi, and Bowen Zhou. 2023. Trustworthy ai: From principles to practices. *ACM Computing Surveys* 55, 9 (2023), 1–46.

[60] Nut Limsopatham and Nigel Collier. 2016. Normalising medical concepts in social media texts by learning semantic representation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1014–1023.

[61] Carolyn E. Lipscomb. 2000. Medical subject headings (MeSH). *Bulletin of the Medical Library Association* 88, 3 (2000), 265.

[62] Dugang Liu, Pengxiang Cheng, Zinan Lin, Xiaolian Zhang, Zhenhua Dong, Rui Zhang, Xiuqiang He, Weike Pan, and Zhong Ming. 2023. Bounding system-induced biases in recommender systems with a randomized dataset. *ACM Trans. Inf. Syst.* 41, 4, Article 108 (October 2023), 26 pages. https://doi.org/10.1145/3582002

[63] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. arXiv:1907.11692. Retrieved from https://arxiv.org/abs/1907.11692

[64] Zhongzhou Liu, Yuan Fang, and Min Wu. 2023. Mitigating popularity bias for users and items with fairness-centric adaptive recommendation. *ACM Trans. Inf. Syst.* 41, 3, Article 55 (July 2023), 27 pages. https://doi.org/10.1145/3564286

[65] Kristian Lum, Yunfeng Zhang, and Amanda Bower. 2022. De-biasing "bias" measurement. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 379–389.

[66] Jing Ma, Ruocheng Guo, Mengting Wan, Longqi Yang, Aidong Zhang, and Jundong Li. 2022. Learning fair node representations with graph counterfactual fairness. In *Proceedings of the 15th ACM International Conference on Web Search and Data Mining*. 695–703.

[67] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. 2018. Learning adversarially fair and transferable representations. In *Proceedings of the International Conference on Machine Learning*. PMLR, 3384–3393.

[68] Masoud Mansoury, Himan Abdollahpouri, Mykola Pechenizkiy, Bamshad Mobasher, and Robin Burke. 2022. A graph-based approach for mitigating multi-sided exposure bias in recommender systems. *ACM Transactions on Information Systems* 40, 2, Article 32 (2022), 31 pages. DOI: https://doi.org/10.1145/3470948

[69] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys* 54, 6 (2021), 1–35.

[70] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems* 26 (2013).

[71] L. Morse, M. H. M. Teodorescu, Y. Awwad, et al. 2022. Do the ends justify the means? Variation in the distributive and procedural fairness of machine learning algorithms. *J. Bus Ethics* 181 (2022), 1083–1095. https://doi.org/10.1007/s10551-021-04939-5

[72] Arvind Narayanan. 2018. Translation tutorial: 21 fairness definitions and their politics. In *Proceedings of the Conference on Fairness, Accountability, and Transparency, New York, USA*. 3.

[73] Richard G. Netemeyer, David G. Dobolyi, Ahmed Abbasi, Gari Clifford, and Herman Taylor. 2020. Health literacy, health numeracy, and trust in doctor: Effects on key patient health outcomes. *Journal of Consumer Affairs* 54, 1 (2020), 3–42.

[74] A. Ng. 2011. Advice for applying machine learning. Stanford Univ., Stanford, CA, USA, Tech. Rep., 2011. [Online]. Available: http://cs229.stanford.edu/materials/ML-advice.pdf

[75] Harrie Oosterhuis. 2023. Doubly robust estimation for correcting position bias in click feedback for unbiased learning to rank. *ACM Trans. Inf. Syst.* 41, 3, Article 61 (July 2023), 33 pages. https://doi.org/10.1145/3569453

[76] Aditya Pal, F. Maxwell Harper, and Joseph A. Konstan. 2012. Exploring question selection bias to identify experts and potential experts in community question answering. *ACM Transactions on Information Systems* 30, 2, Article 10 (2012), 28 pages. DOI: https://doi.org/10.1145/2180868.2180872

[77] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 1532–1543.

[78] Dana Pessach and Erez Shmueli. 2022. A review on fairness in machine learning. *ACM Computing Surveys* 55, 3 (2022), 1–44.

[79] Foster Provost and Tom Fawcett. 2013. *Data Science for Business: What You Need to Know About Data Mining and Data-analytic Thinking*. O'Reilly Media, Inc.

[80] Bogdana Rakova, Jingying Yang, Henriette Cramer, and Rumman Chowdhury. 2021. Where responsible AI meets reality: Practitioner perspectives on enablers for shifting organizational practices. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–23.

[81] Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Xiaojiang Chen, and Xin Wang. 2021. A comprehensive survey of neural architecture search: Challenges and solutions. *ACM Computing Surveys* 54, 4 (2021), 1–34.

[82] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1, 5 (2019), 206–215.

[83] Tetsuya Sakai, Jin Young Kim, and Inho Kang. 2023. A versatile framework for evaluating ranked lists in terms of group fairness and relevance. *ACM Trans. Inf. Syst.* 42, 1, Article 11 (January 2024), 36 pages. https://doi.org/10.1145/3589763

[84] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 1668–1678.

[85] Deven Santosh Shah, H. Andrew Schwartz, and Dirk Hovy. 2020. Predictive biases in natural language processing models: A conceptual framework and overview. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 5248–5264. DOI: https://doi.org/10.18653/v1/2020.acl-main.468

[86] Galit Shmueli. 2010. To explain or to predict? *Statist. Sci.* 25, 3 (2010), 289–310. https://doi.org/10.1214/10-STS330

[87] Galit Shmueli and Otto Koppius. 2011. Predictive analytics in information systems research. *Management Information Systems Quarterly* 35, 3 (2011), 553–572.

[88] Herbert A. Simon. 1998. The science of design: Creating the artificial. *Design Issues* 4, 1/2 (1988), 67–82. https://doi.org/10.2307/1511391

[89] Sriram Somanchi, Ahmed Abbasi, Ken Kelley, David Dobolyi, and Ted Tao Yuan. 2023. Examining user heterogeneity in digital experiments. *ACM Trans. Inf. Syst.* 41, 4, Article 100 (October 2023), 34 pages. https://doi.org/10.1145/3578931

[90] Ryan Steed, Swetasudha Panda, Ari Kobren, and Michael Wick. 2022. Upstream mitigation is not all you need: Testing the bias transfer hypothesis in pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 3524–3542.

[91] Shivashankar Subramanian, Xudong Han, Timothy Baldwin, Trevor Cohn, and Lea Frermann. 2021. Evaluating debiasing techniques for intersectional biases. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 2492–2498.

[92] Yi Chern Tan and L. Elisa Celis. 2019. Assessing social and intersectional biases in contextualized word representations. *Advances in Neural Information Processing Systems* 32 (2019).

[93] Yla R. Tausczik and James W. Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology* 29, 1 (2010), 24–54.

[94] Mike H. M. Teodorescu, et al. 2021. Failures of fairness in automation require a deeper understanding of human-ML augmentation. *Management Information Systems Quarterly* 45, 3 (2021), 1483–1500.

[95] Kelsey Urgo and Jaime Arguello. 2022. Understanding the "Pathway" towards a searcher's learning objective. *ACM Transactions on Information Systems* 40, 4 (2022), 1–43.

[96] Yifan Wang, Weizhi Ma, Min Zhang, Yiqun Liu, and Shaoping Ma. 2023. A survey on the fairness of recommender systems. *ACM Trans. Inf. Syst.* 41, 3, Article 52 (July 2023), 43 pages. https://doi.org/10.1145/3547333

[97] Ryen W. White and Eric Horvitz. 2009. Cyberchondria: Studies of the escalation of medical concerns in web search. *ACM Transactions on Information Systems* 27, 4 (2009), 1–37.

[98] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 38–45.

[99] J. M. Wooldridge. 2009. Omitted variable bias: the simple case. *Introductory Econometrics: A Modern Approach. Mason, OH: Cengage Learning*. 89–93.

[100] Haolun Wu, Chen Ma, Bhaskar Mitra, Fernando Diaz, and Xue Liu. 2023. A multi-objective optimization framework for multi-stakeholder fairness-aware recommendation. *ACM Transactions on Information Systems* 41, 2, Article 47 (2023), 29 pages. DOI: https://doi.org/10.1145/3564285

[101] Le Wu, Lei Chen, Pengyang Shao, Richang Hong, Xiting Wang, and Meng Wang. 2021. Learning fair representations for recommendation: A graph-based perspective. In *Proceedings of the Web Conference 2021*. 2198–2208.

[102] Heng Xu and Nan Zhang. 2022. Goal orientation for fair machine learning algorithms (December 12, 2022). Available at SSRN: https://ssrn.com/abstract=4300581

[103] Forest Yang, Mouhamadou Cisse, and Sanmi Koyejo. 2020. Fairness with overlapping groups; A probabilistic perspective. *Advances in Neural Information Processing Systems* 33 (2020), 4067–4078.

[104] Kai Yang, Raymond Y. K. Lau, and Ahmed Abbasi. 2023. Getting personal: A deep learning artifact for text-based measurement of personality. *Information Systems Research* 34, 1 (2023), 194–222.

[105] Tal Yarkoni and Jacob Westfall. 2017. Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science* 12, 6 (2017), 1100–1122.

[106] Han Zhang, Zhicheng Dou, Yutao Zhu, and Ji-Rong Wen. 2023. Contrastive learning for legal judgment prediction. *ACM Trans. Inf. Syst.* 41, 4, Article 113 (October 2023), 25 pages. https://doi.org/10.1145/3580489

[107] N. Zhang and H. Xu. 2024. Fairness of ratemaking for catastrophe insurance: Lessons from machine learning. *Information Systems Research, Forthcoming*.

[108] Z. Zhao et al. 2023. Popularity bias is not always evil: Disentangling benign and harmful bias for recommendation. In *IEEE Transactions on Knowledge and Data Engineering*, 35, 10 (2023), 9920–9931, 1 Oct. 2023. DOI: 10.1109/TKDE.2022.3218994

[109] Fan Zhou, Yuzhou Mao, Liu Yu, Yi Yang, and Ting Zhong. 2023. Causal-debias: Unifying debiasing in pretrained language models and fine-tuning via causal invariant learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Toronto, Canada, 4227–4241. Retrieved from https://aclanthology.org/2023.acl-long.232

Received 30 March 2023; revised 20 November 2023; accepted 3 January 2024