Information Theoretic Framework For Evaluation of Task Level Fairness

Surbhi Rathore
surbhi_rathore@uri.edu
University of Rhode Island
Kingston, Rhode Island, USA

Sarah M Brown
brownsarahm@uri.edu
University of Rhode Island
Kingston, Rhode Island, USA

ABSTRACT

Focusing on model-specific notions of fairness emphasizes iterative improvements no matter how disparate performance may be. We posit that by examining problem formulations before fitting a model, a more radical solution space, including not building, is more readily accessible. We propose an information theoretic framework for examining problem formulations in terms of fairness. We evaluate the framework as a practical tool by applying it to established benchmark datasets in fair machine learning. We find that our heuristics reliably rank problem formulations by the severity of bias in resultant models. These results unify prior results on limitations of fairness interventions and motivate new classes of interventions for algorithmic fairness.

CCS CONCEPTS

• **Mathematics of computing** → *Information theory*; • **Computing methodologies** → *Machine learning*; • **Social and professional topics** → **Computing profession**.

KEYWORDS

fairness, mutual information, problem formulation

ACM Reference Format:
Surbhi Rathore and Sarah M Brown. 2022. Information Theoretic Framework For Evaluation of Task Level Fairness. In *KDD EAI 22*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Designing fairness interventions for machine learning derived AI systems by focusing solely on the outputs of the fit model unnecessarily constrains the solution space and limits possible outcomes to iterative improvements over the status quo. Instead, a holistic approach to mitigating AI harms requires considering when to not build a system and when to radically change the problem formulation[16, 17]. In healthcare settings, both adding more features[7] and changing the target variable used for training [13] have reduced racial disparities. These successes, however, have been *reactive*, we propose a proactive approach instead. Most calls for holistic examination of algorithmic systems come from outside of machine learning, from researchers who study algorithms as socio-technical artifacts. Attempting to use exclusively technical solutions is an established *trap* that researchers fall into when aiming to improve algorithms' fairness [17].

Technically inclined people still need to be well equipped to engage in these decisions. We propose a technical framework to serve as a lens for examining the concrete aspects of a problem while teams building AI systems engage in a reflective process. Specifically, we do not provide a fairness intervention, but instead a tool to be used in early stages of model development. Our framework relies on mutual information as a formalism to capture the goals of holistic analysis in a format technically-compatible with existing practices in data driven problem formulation. Our goal is that the familiar probability based computation and tractable heuristics make this work immediately applicable for practitioners, while our inspiration was rooted in more holistic critiques of algorithmic system building.

Our main contribution is to provide a technical framework that assists in pulling normative decisions earlier in the development cycle. We begin by defining necessary information theoretic concepts in §2 and apply them to machine learning tasks in §3. We translate classical criteria and propose two heuristic checks for fairness at the level of a task in §4 . We demonstrate that our task level definitions agree with performance metric based definitions by applying them on two benchmark datasets that support multiple problem formulations in §5.

2 INFORMATION THEORY BACKGROUND

We propose using information theory as formalizer [1] for the decisions made during machine learning problem formulation. Information theory is a useful framework for this objective because it does not require the choice of a specific model while providing concreteness in helpful ways for organizing thinking and making assumptions plain. We propose that this framework can serve as a useful translation tool for practitioners and computational researchers who are more accustomed to working with formalisms than the ambiguity of a qualitative assessment while supporting a step-back approach.

Information theory has been used for feature selection, which also occurs prior to fitting a model, and algorithmic fairness interventions. Information theory unifies many feature selection criteria into a single general form [5]. Specifically for fairness, prior work has attempted fair feature selection by quantifying the potential impact of each feature in information theoretic terms [10]. Other works use information theoretic quantities as loss functions to obtain fair classifiers[2] and regression models[18]. Instead of

KDD EAI '22, August 15, 2022, Washington DC

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

producing an optimization problem, we aim to develop heuristic calculations that can aid practitioners in a more reflective design process before fitting a model as usual.

We begin by defining information before introducing mutual information that our definitions rely on. For each quantity, we provide the formal definition and examples of its behavior for how to interpret it's meaning. We will rely on these standard quantities throughout the remainder of the paper.

Information, I, is how much knowing a random variable's value reveals over knowing its distribution. Information is the opposite of randomness in a variable or, its entropy, HȦ constant variable has zero information; once we know its distribution, we know the value. Formally, for a discrete random variable X with values x_i[1]:

$$\mathrm{I}(X) = \sum_i p(x_i) \log p(x_i) = -\mathrm{H}(X) \tag{1}$$

The units of information depend on the type of logarithm used: bits for base two; nats for base e (natural logarithm).

Mutual Information (MI) between two variables is how much one variable tells about the other is written as $\mathrm{I}(X;Y)$. It is often used to measure agreement between categorical quantities, for example to compare two clustering solutions. MI is zero for independent variables and maximal between a variable and itself. Mutual information can also be expressed as Kullback-Leibler divergence between the joint distribution of a pair of variables actual and the product of their marginals–what their joint would be if they were independent. Rewriting MI this way allows us to interpret it as the measurement of how far two variables are from being independent. For two categorical variables, X and Y.

$$\mathrm{I}(X;Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} = \mathrm{D}_{KL}\left(p(X,Y)||p(X)p(Y)\right) \tag{2}$$

Conditional mutual information is the mutual information between two variables after conditioning on a third. We will be most interested in understanding the relation between X and Y given a third variable Z followed by checks on sub categories of Z for task formulations based on conditional mutual information.

$$\mathrm{I}(X;Y|Z) = \sum_z p(z) \sum_{x,y} p(x,y|z) \log \frac{p(x,y|z)}{p(x|z)p(y|z)} \tag{3}$$

3 MACHINE LEARNING TASKS

While tasks are central to all machine learning, they are rarely formally defined. Distilling a task into a problem formulation is a sequence of crucial decisions that are both technical and normative [14]. Here, we formalize these decisions so that we can apply fairness checks during this process. Mentions of task are often limited to multi-task learning, where a task refers to a single prediction target [6], we generalize slightly from this.

We define an *idealized task* to be the prediction of given target Z. This target does not need to be realizable or directly measurable, it can be a *latent* concept that we wish to predict. An idealized task is then *plausible* if there exists a set of signals, X^* that are at least theoretically measurable that could be used for detection.

In most applications, Z is a latent construct, a quantity with conceptual meaning or value, but that is not readily measurable. Instead, a proxy target, Y that is measurable stands in for the true target variable, Z. Proxy targets are also well suited to infeasible if not impossible to measure target. For example, in COMPAS, the goal, or idealized task is to determine if a person will commit another crime. Theoretically, committing a crime Z is a sufficiently macro-level event that an oracle with complete visibility to all actions a person takes could know. However, in a society with any degree of privacy, not every crime that occurs will be recorded. Instead, we use arrest, which is a recorded event, as a proxy, Y.

Additionally, while Z or Y may be predictable from some measurable features, X^*, these may or may not actually be available for training. For example, a company that builds an app for tracking bike rides and hikes, may wish to predict who is interested in working at the company in order to show ads for their job openings. They decide they want to recruit people who are interested in biking, hiking, and other outdoor hobbies, in addition to having the skills for each role (X^*). However, the platforms they will post their ads on will not have a list of people's hobbies; they have web browsing history (X), assume this captures the relevant signal. This reveals another choice in problem formulation: X is not fixed, or predetermined by the goal, it is a choice. A *realized task* is defined by available features, X and a proxy target, Y.

4 FAIRNESS IN REALIZED TASKS

We focus here on the fairness of realized tasks because that is ultimately what will be used to train a model. Fairness in a task considers the possibility for non discrimination, given that an optimal classifier is constructed. This is more general than the formulations for nondiscrimination of a classifier. Classification based criteria seek to balance error rates in some form across groups. In assessing the task, we instead consider how possible it is to learn the relevant relationship across groups.

These criteria enable considering fairness rigorously early in the machine learning product life cycle. They go beyond checking for representative sampling, which is a common data-level check. These definitions allow a practitioner to consider the potential for equity prior to knowing if the model is accurate or not. It is easier to abandon a task when the model does no yet exist, so it is important to consider broad impacts early. These criteria may be applied to decide to train a model or not or to choose among potential training setups.

We translate classical criteria for nondiscrimination and then provide two more relaxations of these that serve as practical heuristics for evaluating problem formulations. Groupwise fairness definitions typically rely on a protected attribute, often constrained to binary for mathematical convenience. We use a demographic variable, A, similarly, however, allow it to be multivariate and continuous. This way, A can incorporate the many identities that a person has, including legally protected attributes such as gender and race. To test for group disparity, some definitions require a discrete, but never binary, A. Ultimately, our goal is to provide tools for a model builder to evaluate the relative fairness of different

[1] information can be defined similarly for continuous variables, with integrals instead of sums.

problem formulations under these criteria to bring more structure into a process that is typically done ad hoc [14].

4.1 Classical Fairness Criteria in Tasks

While a plethora of fairness definitions have been proposed, each fits in one of three broad criteria: independence, separation, and sufficiency [4][2]. These criteria are defined relative to a prediction for each sample $\hat{Y}$, and a protected attribute, A and a ground truth typically denoted by Y[3]. **Independence** requires independence between predictions and protected attribute, $\mathrm{I}(A;\hat{Y}) = 0$; **Separation** conditions on the truth first, $\mathrm{I}(A;\hat{Y}|Y) = 0$; **Sufficiency** conditions on the prediction and requires independence between the truth and demographics, $\mathrm{I}(Y;A|\hat{Y}) = 0$.

We translate these criteria to tasks by replacing the $\hat{Y}$ with the X in each of the prediction-based criteria. This is motivated by assuming that the predictor, $\hat{Y}$, is optimal in the sense that $\mathrm{I}(\hat{Y};Y) = \mathrm{I}(X;Y)$. In practice, this will not be true, but many learning algorithms can be interpreted as minimizing this error.

The independence criterion is the most complex to translate, yielding three variations depending on the translation approach. Using our translation scheme above we find the *problem-independence criterion for tasks*: $\mathrm{I}(X;A) = 0$. Data often reveals information about protected attributes and this strategy, of removing all features related to demographics does not often succeed or leaves an empty set [9]. If we instead focus on the idealized task, we find the *idealized task independence criterion*: $\mathrm{I}(Z;A) = 0$. This is not testable quantitatively, so we can approximate it with the *proxy independence criterion* for tasks: $\mathrm{I}(Y;A) = 0$.

The separation condition for realized tasks is $\mathrm{I}(A;X|Y) = 0$. Consider a case where there is a strong relationship between the features, X and a binary proxy target, Y (which is desirable), for example, two highly separable Gaussian distributed classes. If the target outcome is less prevalent in some groups than others, we would have $\mathrm{I}(X;A) > \mathrm{I}(A;X|Y) = 0$.

The sufficiency criterion for realized tasks requires that $\mathrm{I}(Y;A|X) = 0$. For high dimensional and/or continuous valued features which are common, this becomes much harder to estimate then the above. Especially because in practice, data are sampled in ways that can control and ensure good coverage of the range of the proxy target and perhaps a number of (measured as) discrete demographic variables, but rarely to ensure good coverage of the full support of the feature distribution, without which this calculation may be intractable. The motivation for sufficiency for model fairness is generally rooted in the idea that the prediction is based on features collected in time before the target proxy, making this conditioning more appropriate. In that case, the additional computational complexity may not be a challenge worth resolving, but in a task it may not be.

While none of these criteria will be exactly true, we could aim to choose the problem formulation that is the lowest on our desired quantity or that it must be below a threshold. Minimizing the criterion works especially well for using the separation criterion in order to choose a proxy target. However, establishing a suitable limit in absolute units requires making a judgement of what the acceptable risk is in absolute terms which is poorly defined. Instead, in the following, we will consider comparative terms.

[2]See Chapter 2 of [4] for metrics in each group

[3]While fairness literature refers to Y as ground truth, because it is observed, it is the same as the proxy target we define

4.2 Informative Beyond Demographics

A problem formulation is informative beyond demographics if the data provide more information about proxy target than the demographics, as in Equation 4. This definition approximates problem independence, using the problem of interest, $\mathrm{I}(X;Y)$ to limit the undesirable MI between the features and demographics, $\mathrm{I}(X;A)$.

$$\mathrm{I}(X;Y) > \mathrm{I}(X;A) >= 0 \tag{4}$$

A practitioner should prefer problem formulations with a large gap. We can further consider if this holds equally well for all groups, by conditioning the problem on a group, $A = a$ as in 5.

$$\forall a \in A \qquad \mathrm{I}(X;Y|A=a) > \mathrm{I}(X;A) \tag{5}$$

This task level criterion shares motivation with adversarial approaches to fairness: a fair model predicts the target better than demographic variables. For example, [19] simultaneously trains a classifier to predict the proxy variable well and the demographic variables poorly.

4.3 Equality of Information

A problem formulation passes the equality of information test if the data provides equal information for all demographic groups: $\mathrm{I}(X;Y|A) = \mathrm{I}(X;Y)$. This is a hard constraint, so we introduce a relaxation. A realized task is ϵ-equally informative if for all a_i, a_j

$$|\,\mathrm{I}(X;Y|A=a_i) - \mathrm{I}(X;Y|A=a_j)| < \epsilon. \tag{6}$$

A task being ϵ-fair does not guarantee a fair classifier of any particular hypothesis class exists.

An optimal classifier that passes this test would provide equal accuracy if the classifier itself did not amplify disparities. This test is not sufficient for guaranteeing equal accuracy, however, because of interactions between the learning algorithm's assumption and the data. In a balanced Y test set, this would also approximate independence.

Note that this check does not require that the same subset of features are informative for one group or another or that the two groups share the same decision boundary. A task could pass this check and condition and still require great care in how a predictive model is realized. For example, the two groups could require completely different subsets of the features or different boundaries. In order to realize a predictor that performs well for both groups, the demographic data may be required. This check only validates that it is possible to do so.

5 PROBLEM FORMULATION BENCHMARK VALIDATION

We evaluate our task level formulations by examining the common benchmark datasets for Fair Machine Learning for which varied problem formulations are possible. Evaluation with common benchmark datasets aims to relate these formulations to prior work in

fair machine learning and align with knowledge that these datasets are useful for validating techniques to mitigate bias. We note the limitation of benchmark validation of algorithms intended to be deployed in sociotechnical systems [3, 15]. Our goal, however is not to develop algorithms for deployment, it is to demonstrate a tool to be used in developing new problems for algorithm development. In this context, these benchmarks are an appropriate test bed, even if individual benchmarks do not hold up to scrutiny on their objectives. These benchmarks do serve as examples of problems machine learning is applied to. We selected benchmarks based on prior works that demonstrated the impact of problem formulation on fairness outcomes: one focusing on the varied choice of Y [8], the second finding that cost is a poor proxy in a healthcare system audit [13]. We apply our framework to each problem formulation and show that our criteria are able to identify disparities previously identified in model audits, without training a model.

5.1 Histogram Estimation Quality

To apply our definitions from §4, we need to estimate the distributions from the MI definitions in §2. We evaluate the sensitivity of MI quantities in our definitions to binning for a simple histogram approximation to confirm that our results are robust to the approximation parameters.

We used the *Adult* reconstruction data from [8] with age, gender, race, and marital-status as A, continuous-valued income as Y, and other variables as X[4]. We binned hours-per-week, capital-gain, capital-loss, and income into 6,8,6,20; 3,6,4,30; 4,5,4,40; 6,6,6,30; 4,4,4,30; and 10,7,6,20 bins respectively. We discretized 'age' into intervals of 1-20, 21-40 . . . 81-100. Figure 1 shows that the relative value of MI quantities does not vary. This deems histogram approximation sufficient for our purpose; approximation error does not change our conclusions.

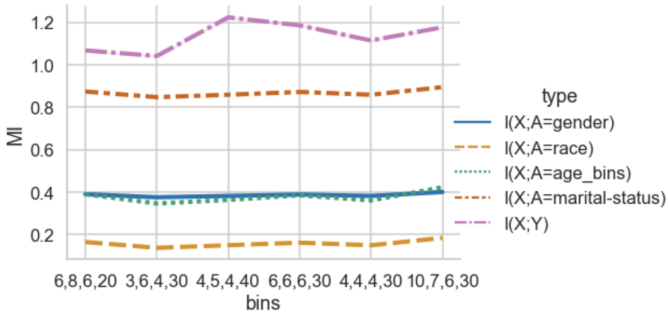

Figure 1: MI for various approximations MI by varying number of bins for numerical features - hours-per-week, capital-gain, capital-loss, income(target). The plot shows variation in absolute MI with varying bin sizes but minimal change in relative value across different heuristics in our formulation.

[4]We did not include protected attributes as features in any of our evaluations, so our formulations have lower MI than what is typically fed to classifiers.

5.2 Comparing Proxy Targets in *Adult*

We test to see that our mutual information based evaluations of the problem formulation replicate the findings of [8] that in a reconstructed version of the adult dataset, different binarizations of income produce varying levels of bias in a naively trained model. In the reconstructed adult dataset the proxy target, income, is continuous, so we constructed 9 separate problems by binarizing the income with thresholds at >=10000, >20000 . . . >=90000. In Figure 2 we see that different formulations have different mutual information, following the results by [8] with different accuracy in the curve for $I(X;Y)$, so our formulation agrees with metrics applied to fit models. Figure 2 also shows that *Adult* fails our Informative Beyond Demographics criterion for protected attributes other than race for all tested candidate proxies. We evaluate the equality of information for each protected attribute in Figure 3, plotting, for each proxy, $I(X;Y|A=a)$. There is some groupwise variation for each protected attribute, but less for gender and it is most extreme for marital status. Crucially this shows that a different threshold on income provides minimum variation for each protected attribute; further complicating the task of selecing a single formulation.

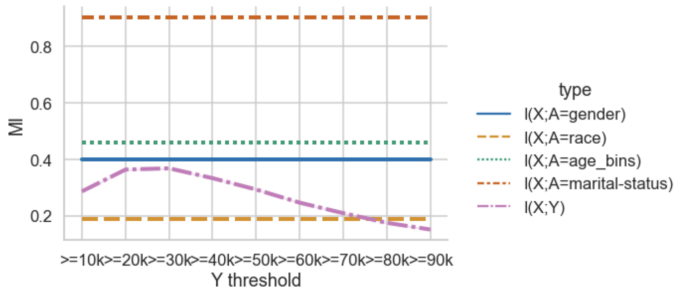

Figure 2: Informative beyond demographics in *Adult*. *Adult* only passes this check for Race with the income threshold set at or above $80k.

5.3 Information Equity of Different Proxy Targets in a Healthcare Risk Score

While the adult reconstruction considers various proxy targets, it is still a fairly artificial problem, we do not know what the target is for that problem. In an audit of a healthcare system's risk score algorithm used to refer patients to managed care Obermeyer et al found using the same features with a different training target decreased racial bias [13]. The released data is matched synthetic data for patient privacy. We used comorbities and biomarkers from a previous year as X, demographic variables as A and compared three values for Y: total cost, avoidable costs, and total number of active chronic illneses. The original findings showed that using cost to train risk models produced one where Black patients had to be much sicker to be enrolled in the managed care program relative to white patients. Obermeyer et al found using instead the total number of chronic illnesses to minimizes the gap. Reducing avoidable expenditures is a goal of the program, so we include that in our analyses as well.

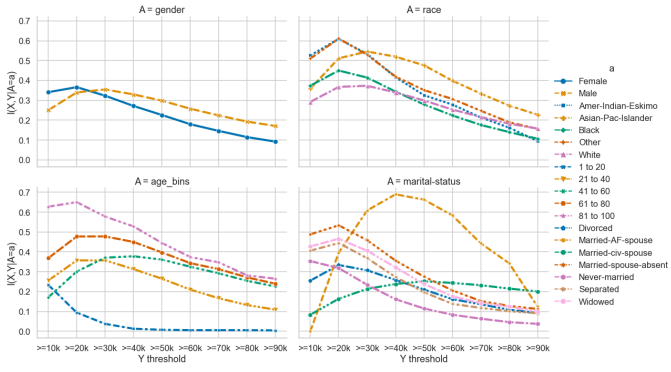

Figure 3: Equality of Information in *Adult*. MI of the Adult features for various binarizations of the income (Y candidates) for each protected attribute in each subplot. For race and gender, different proxies have the same ϵ, but choice of proxy impacts equity relative to marital status.

	avoidable cost	cost	Num active chronic illnesses
Black	2.14945	4.624869	1.868068
white	1.585879	4.143477	1.439384
difference	-0.563571	-0.481392	-0.428684

Table 1: $I(X, Y|A = a)$ for candidate proxy targets by race. The total number of active chronic illnesses provides the smallest information gap.

Table 1 shows that the number of active chronic illnesses provides the smallest information gap. Although, it is the lowest information, this formulation is still informative beyond demographics ($\mathrm{I}(X; A_{\mathtt{race}}) = .271750$). The task information is higher for Black patients than White, but a large gap can still harm Black patients because the population has 10 times as many White patients than Black patients.

6 DISCUSSION AND CONCLUSION

We provide task level fairness criteria that can be applied prior to fitting a model, enabling a model-builder to compare different possible choices for a proxy target or feature set. We envision this work being applied in the early phases of problem formulation and as evidence supporting a choice to not build a model. This set of mutual information based criteria can be used while datasets are being constructed, to evaluate sampling procedures, proxy target selection, and the many other decisions that are made prior to training. Using a low computational cost histogram approximation for distributions of benchmark datasets, we find results consistent with other work that compares problem formulations, without expending the computational cost to train a model.

In tabular data, histogram approximations or kernel density approximations are sufficient and our framework will accompany standard feature engineering procedures. For more complex data, approximation of mutual information with neural networks is a promising direction for broad application [11, 12]. Our derivation provides conceptual relationships(§4.1) and our empirical results provide coarse quantitative verification of the connection between task level metrics and model specific criteria, but detailed exploration of this connection is an open area of work. Extending our translations of classical criteria (§4.1) to feature selection will round out our problem formulation framework.

REFERENCES

[1] Rediet Abebe, Solon Barocas, Jon Kleinberg, Karen Levy, Manish Raghavan, and David G Robinson. Roles for computing in social change. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 252–260, 2020.

[2] Wael Alghamdi, Shahab Asoodeh, Hao Wang, Flavio P Calmon, Dennis Wei, and Karthikeyan Natesan Ramamurthy. Model projection: Theory and applications to fair machine learning. In *2020 IEEE international symposium on information theory (ISIT)*, pages 2711–2716, 2020. tex.organization: IEEE.

[3] Michelle Bao, Angela Zhou, Samantha Zottola, Brian Brubach, Sarah Desmarais, Aaron Horowitz, Kristian Lum, and Suresh Venkatasubramanian. It's COMPASlicated: The messy relationship between RAI datasets and algorithmic fairness benchmarks. In *NeurIPS Dataset Track*, 2021.

[4] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and machine learning*. fairmlbook.org, 2019.

[5] Gavin Brown, Mikel Luj, A Pocock, MJ Zhao, M Luján, and Mikel Luj. Conditional likelihood maximisation: A unifying framework for information theoretic feature selection. *The Journal of Machine Learning Research*, 13:27–66, 2012.

[6] Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997. Publisher: Springer.

[7] Irene Chen, Fredrik D Johansson, and David Sontag. Why is my classifier discriminatory? In *Advances in Neural Information Processing Systems*, pages 3539–3550, 2018.

[8] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. *Advances in Neural Information Processing Systems*, 34, 2021.

[9] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226. ACM, 2012.

[10] Sajad Khodadadian, Mohamed Nafea, AmirEmad Ghassami, and Negar Kiyavash. Information theoretic measures for fairness-aware feature selection. *arXiv preprint arXiv:2106.00772*, 2021.

[11] Valero Laparra, J Emmanuel Johnson, Gustau Camps-Valls, Raul Santos-Rodríguez, and Jesus Malo. Information theory measures via multidimensional Gaussianization. *arXiv preprint arXiv:2010.03807*, 2020.

[12] Qiao Liu, Jiaze Xu, Rui Jiang, and Wing Hung Wong. Density estimation using deep generative neural networks. *Proceedings of the National Academy of Sciences*, 118(15), 2021. Publisher: National Acad Sciences.

[13] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019. tex.publisher: American Association for the Advancement of Science.

[14] Samir Passi and Solon Barocas. Problem formulation and fairness. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 39–48, 2019.

[15] Inioluwa Deborah Raji, Emily M Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. AI and the everything in the whole wide world benchmark. *NeurIPS Dataset Track*, 2021.

[16] Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, pages 33–44, New York, NY, USA, 2020. Association for Computing Machinery. event-place: Barcelona, Spain.

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

[18] Daniel Steinberg, Alistair Reid, Simon O'Callaghan, Finnian Lattimore, Lachlan McCalman, and Tiberio Caetano. Fast fair regression via efficient approximations of mutual information. *arXiv preprint arXiv:2002.06200*, 2020.

[19] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM conference on AI, ethics, and society*, pages 335–340, 2018.

ACKNOWLEDGMENTS

SR was supported by a IBM Global University Program Academic Award to SMB.