How a socio-technical approach to AI auditing can change how we understand and measure fairness in machine learning systems

Gemma Galdon Clavell[1], Ariane Aumaitre[1,2] and Toon Calders[3,†]

[1]*Eticas.ai*
[2]*European University Institute*
[2]*University of Antwerp*

Abstract

Most AI accountability approaches are either too general or too specific. Impact assessments, on the one hand, tend to focus on principles and general commitments, with little or no technical input or transparency. AI auditing tools, on the other, focus on model outputs and predictive accuracy metrics, with no visibility over impacts or structural bias dynamics. We have observed how existing approaches may end up validating unfair and inefficient systems.

In our work at Eticas.ai, we explore the possibilities of a socio-technical approach to AI auditing. We capture key metrics at different stages of a system and put them in relation to relevant social, demographic and sector-specific data. This allows us to assess how protected and other attributes perform at different times in the decision-making process (from training and pre-processing to post-processing), but also in relation to society. Most crucially, our proposal focuses on impact, which is central to all regulatory efforts around the world, and at the heart of impact assessments.

For this specific paper, we show how current approaches to bias measurement are problematic. We use a public dataset, the Adult Census Income Dataset, to show how different bias metrics and approaches lead to different outcomes, and how these may validate unfair AI systems. In the paper we reproduce the criteria used by two well-known bias tools, Fairness 360 and Aequitas, and one regulatory piece, the NYC Bias audit Law, and compare our results with our own approach, which emphasises representativity (as opposed to accuracy or impact in output data) and bias diagnosis. We show how bias metrics that only capture data from model performance and outputs are incomplete and potentially harmful, as they fail to incorporate data from actual, real-world impacts and contextual dynamics such as structural discrimination and power relations. Our auditing methodology overcomes many of these shortcomings and build a robust basis for the development of AI audits as a professional practice.

With this paper and our argument for our "E2E-ST" approach, we wish to open a discussion in the AI bias community and among policy-makers as to how and when to measure impact, and what impact means in AI system performance. Our main claim is that AI bias is not a technical problem with technical solutions, but a socio-technical problem that requires socio-technical approaches to be addressed and lead to the effective protection and fair treatment of discriminated groups and outliers.

Keywords

Responsible AI, Bias audit, Fair machine learning

AIMMES 2024 Workshop on AI bias: Measurements, Mitigation, Explanation Strategies | co-located with EU Fairness Cluster Conference 2024, Amsterdam, Netherlands

*Corresponding author.

gemma@eticas.ai (G. G. Clavell); ariane.aumaitre@eticas.ai (A. Aumaitre); toon.calders@uantwerpen.be (T. Calders)
0000-0002-6481-0833 (G. G. Clavell); 0000-0003-4631-9984 (A. Aumaitre); 0000-0002-4943-6978 (T. Calders)

1. Introduction

The increasing use of personal data to make automated decisions has led to a generalised concern over bias and, specifically, over how AI systems capture and reproduce bias when making decisions. Scholars and researchers have explored various dimensions of bias, aiming to understand its origins, manifestations, and implications across different domains. Contributions from Timnit Gebru [1], Joy Buolamwini [2], Cathy O'Neil [3], Safiya Umoja Noble [4], Ruha Benjamin [5] or Latanya Sweeney [6] have raised awareness on this issue and provided specific examples of AI bias discriminating against women, people of colour, non-English speakers, people with disabilities, and people from specific geographical areas. Overall, the literature shows a continued problem with bias in AI systems, and a clear amplification challenge: data systems may not only be capturing but also amplifying existing, human bias.

The bias problem in AI has found echo in current regulatory debates and developments. Most existing regulation and best practice documents on AI capture the need to conduct risk assessments that look at bias and inclusivity in AI development. So far, the most specific achievement of the AI bias community in terms of turning metrics into regulatory requirements is Law 144 of 2021, also known as the New York City's Bias Audit Law [7], effective since January 1, 2023. It prohibits employers from using automated employment decision tools to evaluate applicants unless the employer meets several requirements prior to its use, including conducting independent audits of such systems, performed by an independent third-party, provide notice of their use and publish a summarized version of the audit/s. Interestingly, NYC's law focuses on a specific bias metric: disparate impact (DI).

This issue has also raised interest in the technical field as well, and conferences such as Fairness, Accountability, and Transparency in Machine Learning (FAccT) are a clear example of the increasing attempts to address this issue and come up with solutions. This attention and awareness of the bias problem in AI has also prompted multiple actors to produce tools to help practitioners measure and address bias. These include:

- IBM AI Fairness 360 (AIF360) [8], an open-source toolkit that provides a set of algorithms, metrics, and bias mitigation techniques for assessing and addressing bias in machine learning models. It includes capabilities for measuring bias in datasets and models, as well as tools for applying pre-processing and post-processing techniques to enhance fairness.
- Aequitas [9]: another open-source bias audit toolkit that allows users to assess bias in machine learning models and datasets. It provides statistical and visual analyses to identify and understand disparities in model predictions across different demographic groups. Aequitas is designed to be customizable and works with various fairness metrics.
- TensorFlow Model Analysis [10], which includes fairness indicators and Mtools for computing fairness metrics and visualizing fairness disparities in machine learning models. The toolkit supports metrics such as disparate impact, equalized odds, and more. It enables users to monitor and evaluate model fairness during development.

This list is far from comprehensive. The tools available can be counted in the dozens. But for all this apparent diversity, existing and available tools are focused on measuring fairness in model performance. As we will show below, this is highly problematic.

We are not the first to point to the need to explore bias beyond model performance. Many critical voices [11] have pointed to the need to complement existing model-centred approaches with further checks, with a particular focus on the need to ensure that end-users are involved in the training and testing of AI systems. In fact, all authors mentioned above have contributed with their work to shaping the debate on the ethical and social dimensions of AI, emphasizing the significance of including end users in the design, testing, and deployment of AI systems and calling for a more inclusive and user-centred approach to AI development.

Our work, however, shows that current approaches to AI bias are fundamentally flawed. As technical and non-technical approaches to bias have developed independently of each other, current socio-technical approaches resemble a sort of methodological Frankenstein where quantitative and qualitative approaches coexist without dialogue or interaction, and the overall picture on how systems impact on people and societies is lost.

In our work, we look at bias from a truly socio-technical perspective and a focus on what we believe (and propose) must be the main entry point for any attempt to make AI accountable: impact. Indeed, all bias metrics we have seen limit their analysis to model data. In the best cases, training data may be taken into account. Therefore, what they all measure is model outputs, not actual societal impacts. This opens a dangerous door to auditing and accountability exercises that validate systems with unfair, problematic or otherwise undesirable impacts. What is the contribution of checking fairness in the limited world of and AI model, when there is no relation between model fairness metrics and actual impact on the world where the AI model will be run? A focus on model metrics and output is also blind to the impact of structural discrimination, power relations and barriers to access, which should be at the core of any AI inspection mechanism.

Drawing on the NYC bias law mentioned above, we found ourselves with a crucial problem: as AI auditors, can we validate a system that complies with DI metrics, where an employer is deemed to be fair when employing 50 of each 100 male candidates and 1 of each 2 female candidates? Ignoring structural discrimination and barriers leads to systems where perpetuating bias is not only seen as acceptable, but also “fair”.

To address these shortcomings and contribute to the development of a robust AI auditing ecosystem, what we propose is an auditing process that captures key model metrics (in-processing), as well as training and labelling data (with a focus on protected attributes, pre-processing). Additionally, having worked extensively in Europe, where AI decisions require a “human in the loop”, we have also come across an additional problem: what if we can assess the predictive accuracy of a system, but have no access to final, real-life decisions? Could we validate a system based on outputs knowing that a series of “humans in the loop” may have altered the initial data?

Our approach is heavily informed by how other sectors have addressed similar challenges. In the medical field, for instance, proving that a solution works in a lab environment is only a stepping stone towards market approval, and many medical products never see the light of day as they fail to demonstrate their benefits and side effects (impacts) in clinical trials. The scalability of impacts is so relevant that clinical trials are organized in phases to ensure that developers adhere to a precautionary principle [12].

Our work on AI auditing at Eticas.ai addresses these questions. As we develop AI auditing software that is socio-technical, we are determined to avoid the methodological Frankenstein and look at systems from the perspective of impact and by using demographic and other relevant data, and not just model data, as our threshold and reference.

2. Existing Fairness Criteria

Existing tools for evaluating model fairness are based on criteria that can be computed based on a dataset containing both the sensitive attributes of the instances, a ground truth label, and the decisions of the model being assessed. In our approach we extend these fairness criteria based on additional demographic and contextual (real-world impact) data to get a more complete picture of the moments and sources of bias in the creation of the model. But, before we dive into the details of our approach, we first revisit the existing criteria as they are used by tools such as AIF and Aequitas, as this helps us build a common ground.

Barocas et al. [13] classify these fairness criteria into 3 groups: those based on Independence, Separation, or Sufficiency:

- Criteria that assess disparity in outcomes. These criteria compare the rate of favourable outcomes for the protected group to the rate of favourable outcomes for the unprotected group. Disparate Impact Ratio (DIR) for instance computes the ratio between them. A DIR value close to 1 indicates fairness, while values significantly different from 1 suggest disparate impact. Similar in nature, Statistical Parity Difference (SPD) measures the difference; an SPD close to 0 indicates fairness, while positive or negative values suggest disparities. This criterion is also known as Demographic Disparity (DD).
- Criteria based on disparity in prediction errors, comparing true positive rates (TPR) and false positive rates (FPR) between protected and unprotected groups. A high difference in TPR indicates that the group with the higher TPR is advantaged as more members that should receive the positive label do so, while a higher FPR indicates that a higher number of members of the group with the higher FPR that should not receive the positive label do receive it nevertheless. The most well-known representant in this category, combining both quantities, is Equalized Odds (EO), which compares the true positive rate and false positive rate between different groups with the following formula: EO=max (TPR protected TPR unprotected, FPR protected FPR unprotected), where EO should be close to 0 for fairness; a higher value indicates disparities in error rates. The Theil index also falls into this category of criteria that are based on disparities in errors; based on the type of error an individual benefit for the receiver of the label is determined, and the Theil index measures how much the distribution of the benefits differs from those in a situation of perfect equality.
- Calibration, which assesses disparities in the interpretation of the predicted labels. For instance, if 70% of instances receiving a positive prediction indeed has the positive label, the interpretation of a positive prediction becomes: 70% chance to be positive. Similarly, the interpretation of the negative label could be: 20% chance to be positive. Calibration-based fairness criteria require these quantities to be the same for the protected and unprotected group. Hence, the interpretation of what it means in terms of true probability of having a positive label given the label assigned by the classifier should not depend on the protected or unprotected group one is in.

Besides these criteria, there are also relaxations that do not require them to hold in general, but only in subgroups of the data. If for instance the protected group overall has fewer financial means than the unprotected group, demanding demographic parity in a loan application may be too strict of a condition. However, within subgroups of people with similar financial means, we would expect similar outcomes and hence demographic parity to hold. Conditional fairness criteria such as Conditional Demographic Disparity (CDD) take this relaxation into account by for instance averaging demographic parity differences over these conditional subgroups.

There are also individual fairness criteria that quantify to what extent an instance in a protected group was treated fairly, by comparing the treatment of that instance with similar instances in reference groups. These individual fairness measures could be used to explore specific cases but are hard to generalize to a fairness audit aimed at assessing a model or automated decision process.

3. The Eticas.ai Approach: from output to impact

We propose an end-to-end, socio-technical approach ("E2E-ST") that considers not only the model, but also the different phases in the model construction and deployment. Whereas other tools typically only consider a single dataset with ground-truth labels and model predictions, we consider the training data that was used, reference data, and operational data. In this way a more complete picture can be made, analysing not only the model's predictions, but also how it was constructed while putting it into the more global context with reference data of the community in which the model is deployed. This approach allows us to produce a rich screen of the model under scrutiny, containing a wide variety of probes and signals.

The architecture of our E2E-ST approach is depicted in Figure 1. Starting from three data sources connected respectively to context, training, and deployment of the model under scrutiny, we first compute statistics called probes. The probes are roughly comparable to measures such as demographic parity difference listed in Section 2. These measures are then logically grouped into signals, that is higher-level indicators of potential issues. For instance, a signal could combine several probes measuring similar properties of the datasets. Based on the signals, a final diagnosis including potential moments of bias can be made. We will now detail the different steps.

3.1. Data used by Eticas.ai's E2E-ST approach

Our approach relies on several datasets provided by the client/system owner:

- The data used to train the model, called training data. This dataset includes the features on which the model bases its decisions, the sensitive attributes indicating membership of protected groups, the ground-truth label, and the predicted label. We assume the predicted label was generated using a methodology that avoids data leakage, such as cross-validation. The predicted label and the ground-truth label allow hence to estimate the accuracy of the model.
- During deployment, data is collected in what we call the "operational dataset". This dataset contains the features on which the model bases its decisions, and the predicted

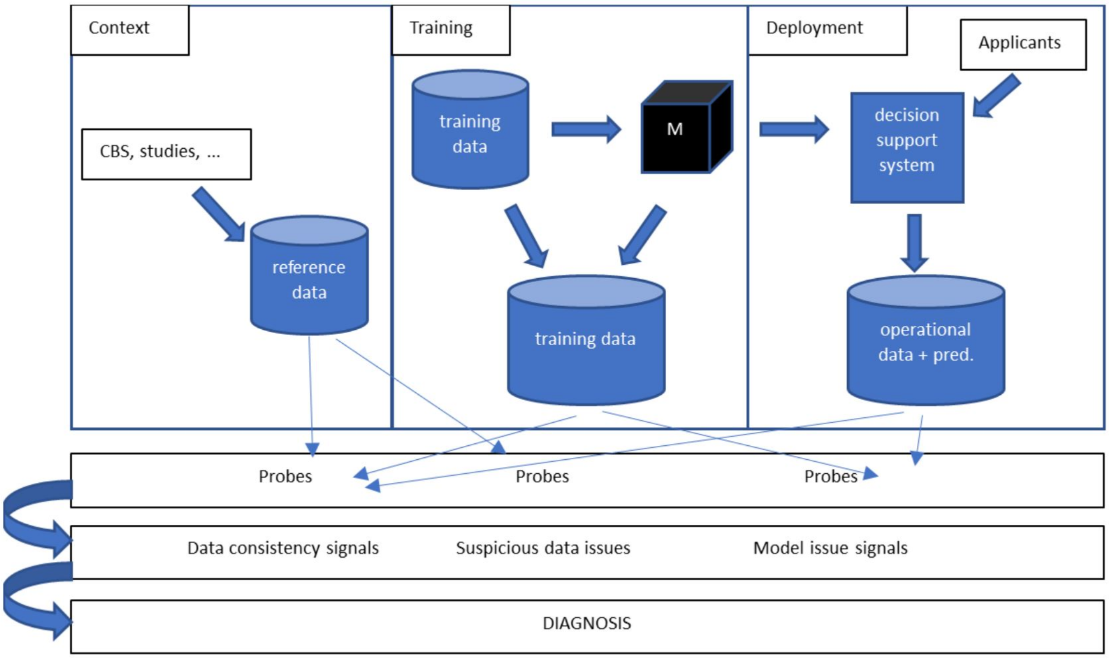

Figure 1: The Eticas.ai Approach

label. We assume that it is uncommon to collect the sensitive data during the operational phase. Also, the ground truth label is not available in the operational dataset.
- Next to training and operational data, we also consider "reference data" that reflects the distribution of the protected groups in the whole population. This dataset can be collected from reliable data sources such as a central bureau of statistics.

We denote the non-sensitive attributes as X_1, …, X_m, the sensitive attributes as S_1, …, S_k, and the ground truth label as Y. A model M is trained and produces a decision D based on X_1, …, X_m.

- The training dataset contains X_1, …, X_m, S_1, …, S_k, Y, D.
- The operational dataset contains attributes X_1, …, X_m, D.
- The reference dataset contains attributes S_1, …, S_k.

3.2. Probes and Signals reported in our method

Based on the available data, Eticas.ai tests for different signals. The signals are tested using several probes, that is, statistics computed from the data. By combining different signals, a coherent story is built of the data, pinpointing potential issues of the data in the form of types of bias that potentially affected the data and led to bias.

3.2.1. Signals for data inconsistencies

The first set of signals tests for inconsistencies between the three data sources. Such consistencies could be indicative for several types of bias, including measurement bias, sample bias, and access bias.

Signals:

- Training distribution shift: the distribution of the training data differs from the distribution of the reference data.
- Operational distribution shift: The distribution of the operational data differs from the distribution of the reference data.
- Training-operation inconsistency: the distribution of the training data differs from that of the operational data.

These signals are based upon statistical tests of differences between distributions. For the third one, a non-parametric test is used based on the performance of a classifier separating training and operational data.

3.2.2. Signals for suspicious data conditions

The second class of signals tests for specific conditions in the data that give rise to potential model problems. For instance, correlations between sensitive data and the ground-truth label may give rise to models picking up or even amplifying historical bias. Another signal is the existence of proxies, or the sensitive attributes having predictive power.

Signals:

- Existence of proxies: there are attributes that can serve as a proxy for the sensitive attributes.
- Training label disparity: the distribution of the given label in the training data differs between sensitive groups.
- Informative sensitive data: the sensitive data contains information that helps to predict the label.

These signals are based upon probes computed on the training data.

3.2.3. Signals for suspicious model conditions

The third class of signals are directly related to model performance.

Signals:

- Poor performance: the predictive performance is low.
- Prediction disparity: the distribution of the predicted label differs between the sensitive groups

- Predictive error disparity: the distribution of the errors differs between the sensitive groups
- Non-calibration: the meaning of the label differs for the different sensitive groups.

These signals roughly correspond to the most used existing fairness measures. As we will argue, however, the interpretation of these signals in absence of the other signals is not possible. Without context these signals are meaningless.

3.3. From Signals to Diagnostics

The third and last step of our E2E-ST audit is the interpretation to come to a diagnosis. Consider the following, entirely fictional example:

Example 1: When screening potential job candidates, a recruiter considers the appearance of a candidate (professional attire) when deciding. The appraisal of the appearance, however, is not recorded. Later, based on historical data generated by the hiring process, a model is trained to automate the candidate selection process. Suppose now that the candidates' attire was more likely to positively influence the decision for female candidates. In the presence of proxies for gender, a model will likely pick that up and exploit correlations between the proxies and the label, leading to a biased model because of an omitted variable bias.

In this case, there will be a combination of signals that together can be interpreted to come to the right diagnosis. For instance, there will be the signals: existence of proxies, training label disparity, prediction disparity, and likely predictive error disparity.

Example 2: The same setting, but now the recruiter does not consider attire. The company, however, has a bad reputation of not hiring females. Because of this reputation, only strong female candidates apply. Furthermore, being aware of its bad reputation, the company starts monitoring the acceptance rates, actively encouraging equal acceptance rates between females and non-females. Also here, based on historical data generated by the hiring process, a model is trained to automate the candidate selection process. This model will pick up the higher requirements for females, but this will not result in a prediction disparity signal, because of the combination of a self-selection bias and a historical bias.

In this case, the end-to-end, socio-technical approach considering different perspectives, including impact, will raise an alert pointing to training distribution shift and the existence of proxies, for instance.

3.4. Comparative overview

We believe that the main contributions of our approach are, on the one hand, taking context and impact into account, including demographic reference information and operational data. On the other hand, we move beyond identifying probes and signals to diagnosing where the source for bias may be located. The following table summarizes the types of data used by different approaches, showing how our end-to-end, socio-technical auditing approach has visibility over crucial, impact data and identifies bias sources:

	Based on training data	Based on operational data collected during deployment	Using impact, demographic data	Combining different measures in a diagnostic phase to identify the moments of bias
Eticas.ai	×	×	×	×
AI Fairness 360		×		
Aequitas	×			
NYC methodology	×			

4. Eticas.ai vs other approaches: a comparative case study

To show how our end-to-end, socio-technical approach compares to other proposals, we have compared bias detection results in the Adult Census Income dataset using four distinct fairness methodologies: Fairness 360, Aequitas, the NYC methodology, and our E2E-ST method. Our claim is that current tools do not identify bias in impact, nor provide useful workflows for auditing. Fairness 360 and Aequitas are toolboxes that can be used flexibly, they do not include guidelines or clear workflows for auditing. The closest to workflows for auditing are the demos that are available for Fairness 360 [14] and Aequitas [15]. Therefore, we opted to include the results produced by these demos in our analysis. The NYC methodology is named after the description of what an audit needs to include in the NYC local law 144 [16]. This legislation makes audits mandatory for automated employment decision tools. The law states that such a bias audit shall include but not be limited to the testing of an automated employment decision tool to assess the tool's disparate impact on persons of any component 1 category [...]. In our comparison we hence report disparate impact as per the "NYC methodology".

Our analysis aims to evaluate how each methodology identifies and measures bias across two primary sensitive attributes: sex and ethnicity. Furthermore, we adopt an intersectional lens, applying these methodologies to the intersection of sex and ethnicity wherever possible. This approach not only aligns with recent advances in understanding discrimination and bias but also shifts the focus from mere output analysis to assessing real-world impacts. Following the scenario set forth in Section 3.1, we assume that two datasets are available: the training dataset and the operational dataset, and we assume to have reference data at our disposal. As the tools of Aequitas and Fairness 360 both require the model's decisions and the ground-truth label, they can only be applied on the training dataset. On the other hand, the NYC methodology concerns the impact, post-hoc, of existing automated decision support systems, making the operational dataset the most logical one to apply the methodology on.

4.1. The data: the Adult Census Income Dataset

To illustrate our approach, we generated a realistic scenario utilizing data from the Adult Census Income dataset, a prominent dataset employed in machine learning for the purpose of income prediction. This dataset originates from the 1994 US Census database and is esteemed for its comprehensive demographic and employment information. It includes a variety of attributes such as age, education, race, sex, income, and native country, making it an indispensable resource for analytical studies in machine learning, particularly in the evaluation of fairness and bias. The principal predictive target within this dataset is income, which is bifurcated into two categories: individuals earning more than $ 50,000 annually and those earning less.

For the purposes of simplification and to facilitate our analysis, several data transformations were applied:

- The race variable was recoded into two categories: White and Non-White.
- The sex variable was recoded into a binary category: “Female” and “Other”[1]
- Education levels were consolidated into three categories: Low, Medium, and High, to provide a more generalized view of educational attainment.

On this dataset then decision tree models were trained, using 10-fold cross validation. In this way for each data instance a prediction was made based on a model trained on all folds, but the fold the instance was in. This labeled dataset was then split into two parts, one to represent the training data, and the other the operational data. The reference dataset for demographic comparisons was collected from the 2020 Census of the U.S. Census Bureau and uses national averages for gender and ethnicity.

4.2. Methods

To ensure a robust and fair comparison, our analysis maintains consistency in both the dataset and the model used across all methodologies. Recognizing that the potential for bias detection may vary significantly depending on the model applied, we employ the same predictive model across all methodologies.

As a methodological note, it should be mentioned that for certain probes where the dataset is compared against real-world data, the analysis compares demographic data from 2022, while the census data is from 1994. We believe that this is not relevant to our argument in any way that disqualifies our results, but must be mentioned. We are working to further test our approach and will continue to share results and lessons learned with the AI auditing community.

[1] We acknowledge that gender is a spectrum, not a binary construct. The recoding of the sex variable into “Female” and “Other” categories in our dataset transformations is a simplification made for analytical purposes. This approach does not fully capture the diversity of gender identities. Our use of binary classification is constrained by the limitations of the original dataset and is not intended to overlook or diminish the significance of non-binary and transgender identities.

4.3. Comparative analysis

4.3.1. Aequitas—gender

	Training data		Operational data	
	Result	Bias	Result	Bias
Equal Parity	0.17	Failed	N/A	N/A
Proportional Parity	0.35	Failed	N/A	N/A
False Positive Rate Parity	0.33	Failed	N/A	N/A
False Discovery Rate Parity	1.2	Passed		
False Negative Rate Parity	1.12	Passed		
False Omission Rate Parity	0.33	Passed		

4.3.2. Aequitas—ethinicity

	Training data		Operational data	
	Result	Bias	Result	Bias
Equal Parity	0.11	Failed	N/A	N/A
Proportional Parity	0.64	Failed	N/A	N/A
False Positive Rate Parity	0.7	Failed	N/A	N/A
False Discovery Rate Parity	1.25	Passed		
False Negative Rate Parity	1.03	Passed		
False Omission Rate Parity	0.56	Passed		

All criteria computed by Aequitas require the sensitive attribute. In our setting, we assume this information to only be available in the training data.

4.3.3. NYC Methodology

1) GENDER (reference: male)

	Training data		Operational data	
	Result	Bias	Result	Bias
Impact Ratio	0.351	Bias	0.365	Bias

2) RACE (reference: white)

	Training data		Operational data	
	Result	Bias	Result	Bias
Impact Ratio	0.369	Bias	0.614	Bias

3) Intersectional approach

	Operational data	
	Result	Bias
Male, non-white	0.711	Bias
Female, white	0.358	Bias
Female, non-white	0.258	Bias

The NYC methodology requires the sensitive attribute to be recorded in the operational data but could also be executed on the training data. Although we assume the sensitive attributes are not recorded in the operational data, we made an exception for the NYC methodology as it is explicitly requiring this.

4.3.4. AIF 360

1) gender

	Training data		Operational data	
	Result	Bias	Result	Bias
Statistical Parity Difference	-0.158	Bias	N/A	N/A
Equal Opportunity Difference	-0.0522	No bias	N/A	N/A
Average Odds Difference	0.058	No bias	N/A	N/A
Disparate Impact	2.85	Bias	N/A	N/A
Theil Index	0.342	No threshold	N/A	N/A

2) ethnicity

	Training data		Operational data	
	Result	Bias	Result	Bias
Statistical Parity Difference	-0.072	No bias	N/A	N/A
Equal Opportunity Difference	-0.012	No bias	N/A	N/A
Average Odds Difference	0.017	No bias	N/A	N/A
Disparate Impact	1.56	Bias	N/A	N/A
Theil Index	0.342	No threshold	N/A	N/A

4.4. Discussion

Looking at the statistics computed for the model by IBM 360 fairness demo, we observe that the disparate impact (DI) measures for race show deviations that could indicate lack of fairness in the model. The Statistical Parity Difference (SPD), however, does not raise a flag, although both measures look at the difference in acceptance rates between ethnic groups and compare them using a ratio (DI) or the difference (SP). The difference is around -7%, which means that one ethnic group has 7% lower chance of receiving the positive label than the other. The ratio is 1.56%, which indicates that this 7% difference makes the relative difference 156%. The error-based measures show that the errors made by the model show no significant difference between the ethnic groups. The conclusions that can be drawn from these measures, however, are very limited. It is unclear to what extent the disparity in the predictions present an existing disparity in the population, or maybe a historical bias being amplified by the model, or the result of a selection bias where members of one of the groups were more self-selective leading to a higher acceptance rate.

The Aequitas toolkit computes roughly the same statistics as the IBM 360 fairness toolkit demo, although it further splits up the error-based measures in all four categories. The conclusions here are similar as for the IBM360 fairness toolkit, showing a disparity in impact of the model. In the Aequitas tool the negative result for the proportional parity test is deemed to be alarming, stating: If your desired outcome is to intervene proportionally on people from all races, then you care about this criteria. The IBM 360 tool does not make any claims regarding whether there is bias in the model and if we should care about it, but lets the user chose between different mitigation techniques, which is, at the very least, an implicit indication that something needs to be corrected.

There are issues with these error measures, however:

- Suppose that the data has different base rates for different protected communities. Then, no matter what model is presented in the tool, either the statistical disparity of the model will be too high, or there will be a disparity in one of the error-based measures. Indeed: in the original data there is a difference of 19% between males receiving the positive label and females receiving the positive label. The only situation that allows to close this gap is if the way in which predictions diverge from training labels differs between males and females. To close the gap, either many more females with a negative label receive a positive prediction nevertheless than males, or many more males with a positive label receive a negative prediction. Hence, either respectively the FPR or the TPR for females is higher than for males, triggering the other alert. For the FPR criterion Aequitas states that it "is important [to meet the criterion] in cases where your intervention is punitive and has a risk of adverse outcomes for individuals." So, there is no winner here: every model will be deemed unfair.
- The measures are computed on the given data without questioning the data itself. Is the data representative for the population? Or was there selection bias? Are the labels in the training data correct? Are all attributes correctly recorded?

Tools like those described here suffer from their own techno-solutionism bias. Specifically, the belief that the problem can be solved solely by a technological intervention in the form of a set of measures and a pre-programmed mitigation technique making models fair by design.

We propose an alternative, end-to-end, socio-technical approach that considers not only training data and model outcomes on this training data, but also reference distributions of the relevant population, and operational data. This allows for a much broader analysis of the quality of not only the model, but also training data and operational data. Following on our medical analogy, just like a doctor does not look at one measure or a set of measures to assess a patient's health condition, but instead builds up a diagnosis combining different measurements to paint a complete picture, bias assessment and auditing needs to follow a similar approach.

For this case study, the additional tests conducted by the team at Eticas.ai reveal the following additional issues, not reported by the other tools:

- The data is not representative for the general population. Whites and males are significantly over-represented in the training data raising the question of potential sample bias in the data.
- The training data labels themselves show a high disparity, even higher than the disparity of the model. This raises concerns regarding label bias or historical bias; is the data even suitable for learning a model? Or evaluating a model?
- The sensitive attribute gender does not contribute much for the predictions by the model. Given that there is a significant correlation between gender and the label (SPD is a form of correlation between sensitive attributes and labels), this implies that other attributes capture the same information with respect to the label as gender does. Any historical gender-bias in the data could hence be picked up by a model, even if it does not use gender in its predictions.

The combination of all these factors paints a quite worrisome picture: the data does not correctly represent the general population, potentially leading to less qualitative models for underrepresented groups. Additionally, the training data shows a significant label disparity stemming from either label bias (e.g. using a biased proxy for the label) or historical bias. Without further justification of the data explaining the disparities, the training data should be disqualified. Furthermore, the probes indicate a real danger of the models picking up label bias, because gender can be predicted with high accuracy given the other attributes. This is confirmed by both the statistical parity difference in the model's predictions and the fact that the sensitive attribute does not carry additional information on the label.

Overall, we show that most of what has been said about AI fairness, and the existing tools and approaches to bias identification and auditing are very problematic. If we continue to see AI bias as a technical problem with technical solutions, we will validate models that treat people differently for reasons due to their race, gender or other protected attributes.

What we suggest is to mobilise a socio-technical approach to AI inspection that incorporates methods and insights from other fields, and specifically quantitative social sciences. By focusing on impact, we shift the emphasis from the technical to the social, and make AI models and tools accountable not only for their inner workings, but their effects on the societies they draw their data from.

5. End-to-end, socio-technical methods and the future of AI auditing

This is the first part of a long-term project on how to conduct end-to-end, socio-technical AI audits in practice. This initial work on probe comparisons in recommender/ranking systems will be followed by further testing in different verticals. We are building libraries that allow us to quickly establish what is the context data relevant for testing and establishing thresholds –while in this paper we have used general demographic data, we are also testing recommender/ranking systems that may be relevant only to population subsets (patients with specific medical pathologies, for instance). We are also working on adding new probes that add complexity to AI auditing. We aim to further develop our auditing software to test for bias in ways that are robust but also contextual. Our goal is to show that AI auditing can be automated without reducing complexity and context.

This first product has a very specific objective: to open a debate in the AI bias community on how to capture impacts, and not just outputs. Our work and experience show that the AI bias community can benefit from moving away from model-centric approaches and incorporating context data into the metrics and methodologies developed for bias assessment. As we put it earlier, the current focus on model metrics and output leaves out the impact of structural discrimination, power relations and barriers to access, which should be at the core of any AI inspection mechanism if AI auditing is to help developers, policy-makers and society at large better understand and measure how AI systems perform and impact on society.

References

[1] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. D. Iii, K. Crawford, Datasheets for datasets, Communications of the ACM 64 (2021) 86–92.

[2] J. Buolamwini, T. Gebru, Gender shades: Intersectional accuracy disparities in commercial gender classification, in: Conference on fairness, accountability and transparency, PMLR, 2018, pp. 77–91.

[3] C. O'neil, Weapons of math destruction: How big data increases inequality and threatens democracy, Crown, 2017.

[4] S. U. Noble, Algorithms of oppression: How search engines reinforce racism, in: Algorithms of oppression, New York university press, 2018.

[5] R. Benjamin, Race after technology: Abolitionist tools for the new Jim code, Oxford University Press, 2020.

[6] L. Sweeney, Discrimination in online ad delivery, Communications of the ACM 56 (2013) 44–54.

[7] New York City Department of Consumer and Worker Protection, Dcwp-aedt faq, https://www.nyc.gov/assets/dca/downloads/pdf/about/DCWP-AEDT-FAQ.pdf, 2024. Accessed: 2024-03-21.

[8] IBM, Aif 360 github repository, https://github.com/IBM/AIF360, 2024. Accessed: 2024-03-21.

[9] AEQUITAS, Aequitas github repository, https://github.com/dssg/aequitas, 2024. Accessed: 2024-03-21.

[10] Tensorflow, Tensorflow fairness indicators, https://www.tensorflow.org/tfx/guide/fairness_indicators, 2024. Accessed: 2024-03-21.

[11] M. S. Lam, M. L. Gordon, D. Metaxa, J. T. Hancock, J. A. Landay, M. S. Bernstein, End-user audits: A system empowering communities to lead large-scale investigations of harmful

algorithmic behavior, Proceedings of the ACM on Human-Computer Interaction 6 (2022) 1–34.

[12] E. M. Agency, From laboratory to patient: the journey of a medicine assessed by EMA, https://www.ema.europa.eu/en/documents/other/laboratory-patient-journey-centrally-authorised-medicine_en.pdf, 2024. Accessed: 2024-03-21.

[13] S. Barocas, M. Hardt, A. Narayanan, Fairness and Machine Learning: Limitations and Opportunities, MIT Press, 2023.

[14] IBM, Ibm's ai fairness 360 demo, https://aif360.res.ibm.com/data, 2024. Accessed: 2024-03-21.

[15] AEQUITAS, Aequitas' bias report, http://aequitas.dssg.io/, 2024. Accessed: 2024-03-21.

[16] New York City, Local law 144 of 2021: Automated employment decision tools, https://legistar.council.nyc.gov/View.ashx?M=F&ID=10399761&GUID=F99584B7-57C8-469E-9637-46A0E780690E, 2021. Accessed: 2024-03-21.