SUPPORT POOL OF EXPERTS PROGRAMME

AI-Complex Algorithms and effective Data Protection Supervision

Bias evaluation

by Dr. Kris SHRISHAK

As part of the SPE programme, the EDPB may commission contractors to provide reports and tools on specific topics.

The views expressed in the deliverables are those of their authors and they do not necessarily reflect the official position of the EDPB. The EDPB does not guarantee the accuracy of the information included in the deliverables. Neither the EDPB nor any person acting on the EDPB's behalf may be held responsible for any use that may be made of the information contained in the deliverables.

Some excerpts may be redacted or removed from the deliverables as their publication would undermine the protection of legitimate interests, including, inter alia, the privacy and integrity of an individual regarding the protection of personal data in accordance with Regulation (EU) 2018/1725 and/or the commercial interests of a natural or legal person.

TABLE OF CONTENTS

1	Stat	e of the art for bias evaluation5	5
	1.1	Sources of bias5	>
	1.1.	1 Bias from data5	>
	1.1.	2 Algorithm bias6	5
	1.1.	3 Evaluation bias6	5
	1.1.4	4 Sources of bias in facial recognition technology7	1
	1.1.	5 Sources of bias in generative AI7	1
	1.2	Methods to address bias8	3
	1.2.	1 Pre-processing9)
	1.2.	2 In-processing11	L
	1.2.	3 Post-processing	L
	1.2.	4 Methods for generative AI)
2	Тоо	ls for bias evaluation13	3
	2.1	IBM AIF36013	3
	2.2	Fairlearn13	3
	2.3	Holistic AI14	ł
	2.4	Aequitas14	ł
	2.5	What-If Tool14	ł
	2.6	Other tools considered15	;
Conclusion			;
Bibliography			

Document submitted in March 2024

1 STATE OF THE ART FOR BIAS EVALUATION

Artificial intelligence (AI) systems are socio-technical systems whose behaviour and outputs can harm people. Bias in AI systems can harm people in various ways. Bias can result from interconnected factors that may together amplify harms such as discrimination (European Union Agency for Fundamental Rights, 2022; Weerts et al., 2023). Mitigating bias in AI systems is important and identifying the sources of bias is the first step in any bias mitigation strategy.

1.1 Sources of bias

The AI pipeline involves many choices and practices that contribute to biased AI systems. Biased data is just one of the sources of biased AI systems, and understanding its various forms can help to detect and to mitigate the bias. In one application, the lack of representative data might be the source of bias, e.g., medical AI where data from women with heart attacks is less represented than men in the dataset. In another, the proxy variables that embed gender bias might be the problem, e.g., in résumé screening. Increasing the dataset size for women could help in the former case, but not in the latter case.

In addition to bias from data, AI systems can also be biased due to the algorithm and the evaluation. These three sources of bias are discussed next.

1.1.1 Bias from data

- 1. Historical bias: When AI systems are trained on historical data, they often reflect societal bias which are embedded in the dataset. Out-of-date datasets with sensitive attributes and related proxy variables contribute to historical bias. This can be attributed to a combination of factors: how and what data were collected and the labelling of the data, which involves subjectivity and the bias of the labeller. An example of historical bias in AI systems has been shown with word embedding (Garg et al., 2018), which are numerical representations of words and are used in developing text generation AI systems.
- 2. Representation bias: Representation bias is introduced when defining and sampling from the target population during the data collection process. Representation bias can take the form of availability bias and sampling bias.
 - a. Availability bias: Datasets used in developing AI systems should represent the chosen target population. However, datasets are sometimes chosen by virtue of their availability rather than their suitability to the task at hand. Available datasets often underrepresent women and people with disabilities. Furthermore, available datasets are often used out of context for purposes different from their intended purpose (Paullada et al., 2021). This contributes to biased AI systems.
 - b. Sampling bias: It is usually not possible to collect data about the entire target population. Instead, a subset of data points related to the target population is collected, selected and used. This subset or sample should be representative of the target population for it to be relevant and of high quality. For instance, data collected from scraping Reddit or other social media sites are not randomized and are not representative of the population that don't use these sites. Such data are not

generalizable for wider population beyond these sites. And yet, the data are used in AI models deployed in other contexts.

When defining the target population, the subgroups with sensitive characteristics should be considered. An AI system built using a dataset collected from a city will only have a small percentage of certain minority groups, say 5%. If the dataset is used asis, then the outputs of this AI system will be biased against this minority group because they only make up 5% of the dataset and the AI system has relatively less data to learn from about them.

- 3. Measurement bias: Datasets can be the result of measurement bias. Often, the data that is collected is a proxy for the desired data. This proxy data is an oversimplification of the reality. Sometimes the proxy variable itself is wrong. Furthermore, the method of measurement, and consequently, the collection of the data may vary across groups. This variation could be due to easier access to the data from certain groups over others.
- 4. Aggregation bias: False conclusions may be drawn about individuals or small groups when the dataset is drawn from the entire population. The most common form of this bias is Simpson's paradox (Blyth, 1972) where patterns observed in the data for small groups disappear when only the aggregate data over the entire population is considered. The most well-known example of this comes from the UC Berkeley admissions in 1973 (Bickel et al., 1975). Based on the aggregate data, it seemed that women applicants were rejected significantly more than men. However, the analysis of the data at the department level revealed that the rejection rates were higher for men in most departments. The aggregate failed to reveal this because a higher proportion of women applied to departments with low overall acceptance rate than they did to departments with high acceptance rate.

1.1.2 Algorithm bias

Although much of the discussion around bias focusses on the bias from data, other sources of bias that contribute to discriminatory decisions should not be overlooked. In fact, AI models reflect biased outputs not only due to the datasets but also due to the model itself (Hooker, 2021). Even when the datasets are not biased and are properly sampled, the algorithmic choices can contribute to biased decisions. This includes the choice of **objective functions, regularisations, how long the model is trained,** and even the choice of statistically biased estimators (Danks & London, 2017).

The various trade-offs made during the design and development process could result in discriminatory outputs. Such trade-offs can include model size and the choice of privacy protection mechanisms (Ferry et al., 2023; Fioretto et al., 2022; Kulynych et al., 2022). Even with Diversity in Faces (DiF) dataset that has broad coverage of facial images, an AI model trained with certain differential privacy techniques disproportionately degrades performance for darker-skinned faces (Bagdasaryan et al., 2019). Furthermore, techniques to compress AI models can disproportionally affect the performance of AI models for people with underrepresented sensitive attributes (Hooker et al., 2020).

1.1.3 Evaluation bias

The performance of AI systems is evaluated based on many metrics, from accuracy to "fairness". Such assessments are usually performed against a benchmark, or a test dataset. Evaluation bias arises at this stage because the benchmark itself could contribute to bias.

Al systems can perform extremely well against a specific test dataset, and this test performance may fail to translate into real-world performance due to "overfitting" to the test dataset. This is especially a problem if the test dataset carries over historical, representation or measurement bias.

For instance, if the test dataset was collected from the USA, it is unlikely to be representative for the population in Germany; or, if the dataset was collected in 2020 during COVID-19 but used in a medical setting in a non-COVID-19 year. This means, that even if the bias in the training dataset is mitigated, bias might creep in at the evaluation stage.

1.1.4 Sources of bias in facial recognition technology

Historical, representation and evaluation bias are the main causes of bias in facial recognition technology (FRT) and, more broadly, image recognition. This is because the training and benchmark datasets are constructed from publicly-available image datasets, often through web scraping, that are not representative of different groups and different geographies (Raji & Buolamwini, 2019).

Databases such as Open Images and ImageNet mostly contain images from the USA and the UK (Shankar et al., 2017). IJB-A and Adience have been shown to mostly contain images of people with light-skin and underrepresenting people with dark skin (Buolamwini & Gebru, 2018). Furthermore, racial slurs and derogatory phrases get embedded during the labelling process of images (Birhane & Prabhu, 2021; Crawford & Paglen, 2021). And despite datasets being flagged for removal, some of these datasets are still being used (Peng, 2020). If these are used for training and/or testing FRT, then, by design, they'll be biased.

Even datasets that attempt to address the problem can fail in the process. IBM's "Diversity in Faces" dataset was introduced to address the lack of diversity in image datasets (Merler et al., 2019). However, it raised more concerns (Crawford & Paglen, 2021). First, the images were scraped from the website Flickr without the consent of the site users (Salon, 2019). Second, it uses skull shapes as an additional measure, which has historically been used to show racial superiority of white people and, hence, embeds historical bias (Gould, 1996). Finally, the dataset was annotated by three Amazon Turk workers who guessed the age and gender of the images that were scraped.

1.1.5 Sources of bias in generative AI

Generative AI allows for the generation of content including text, images, audio and video. The sources of bias discussed in the previous sections—bias from data, algorithm bias and evaluation bias—get carried over to AI that generates content. In addition, generative AI systems are developed with large amounts on uncurated data scraped from the web. This adds an additional layer of risk as the developers would lack adequate knowledge about the data and its statistical properties, making it harder to assess the sources of bias.

Furthermore, many of the generative AI models are developed without an intended purpose. A pretrained model is built and then applications are developed on top of this pre-trained model by other organisations. Thus, the source of bias can be in the pre-trained model and in the context of the downstream application. When bias is embedded in the pre-trained model, the bias will propagate downstream to all the applications.

Generative AI datasets can reflect historical bias, representation bias and evaluation bias (Bender et al., 2021). Bias can also arise due to data labelling, especially when fine-tuning a pre-trained model for a specific application. Labels or annotations are often added to the data by underpaid workers and Amazon Turks. They may choose the wrong labels because they are distracted, or worse, because they embed their own bias by not being from the representative population where the AI system will be

deployed. This is especially the case when more than one label could potentially apply to the data (Plank et al., 2014).

Although the dataset used for pre-trained model is currently neither curated nor labelled by humans (which organisations claim to be costly), the process of reinforcement learning from human feedback used by companies developing generative AI introduces the same biases, albeit at a later stage in the development process.

Even when the text datasets are well-labelled, they can contain societal bias that arise due to spurious correlations, which are statistical correlations between features and outcomes. In the case of text generative AI, such spurious correlations can be observed with word embeddings, which underlie text generative AI (Garg et al., 2018): e.g., 'man' being associated with 'programming' and 'woman' being associated with 'homemaker'. Furthermore, as these are mathematical objects, the contextual information about the words get lost, and they have been observed to output "doctor" - "father" + "mother" as "nurse." Pre-trained language models such as GPT that rely on uncurated datasets are also susceptible to this issue (Tan & Celis, 2019), and merely increasing the size of the model does not address the problem (Sagawa et al., 2020).

1.2 Methods to address bias

No automated mechanism can fully detect and mitigate bias (Wachter et al., 2020). There are inherent limitations with technical approaches to address bias (Buyl & De Bie, 2024). These approaches are necessary, but not sufficient for AI systems, which are socio-technical systems (Schwartz et al., 2022). The most appropriate approaches depend on the specific context for which the AI system is developed and used. Moreover, the contextual and socio-cultural knowledge should complement these technical approaches.

Based on when the intervention is made in the AI lifecycle to mitigate bias, the technical methods and techniques to address bias can be classified into three types (d'Alessandro et al., 2017):

- 1. Pre-processing: These techniques modify the training data before it is used to train an AI model to obscure the associations between sensitive variables and the output. Pre-processing can help identify historical, measurement and representational bias in data.
- 2. In-processing: These techniques change the way the AI training process is performed to mitigate bias through changes in the objective function or with an additional optimisation constraint.
- 3. Postprocessing: These techniques treat the AI model to be opaque and attempt to mitigate bias after the completion of the training process. The assumption behind these techniques is that it is not possible to modify the training data or the training/learning process to address the bias. Thus, these techniques should be treated as a last resort intervention.

Merely removing sensitive variables from the dataset is not an effective approach to mitigate bias due to the existence of proxy variables (Dwork et al., 2012; Kamiran & Calders, 2012).

Pre-processing approaches are agnostic to the AI type as it focusses on the dataset. This is an important advantage. Furthermore, many of the approaches have been developed and tested over the past decade and are more mature than in-processing techniques. Pre-processing approaches are early-stage intervention and can assist with changing the design and development process. However,

if these techniques are the only intervention used, they might give the illusion that all the bias has been resolved—which is not the case (Obermeyer et al., 2019). they are only the starting point.

For regulators, preprocessing techniques are useful only if they have access to the datasets that were used to train the model. Furthermore, the regulator needs to consider whether other in-processing and post-processing techniques were used by the developer and deployers of the AI system.

1.2.1 Pre-processing

- 1. Data provenance (Cheney et al., 2009; Gebru et al., 2018): Data provenance is an *essential* step before other methods to mitigate bias from data can be used. It attempts to answer where, how and why the dataset came to be, who created it, what it contains, how it will be used, and by whom. In the area of machine learning, the term 'datasheet' is more commonly used. Data provenance can, in the context of data protection, include the listing of personal data and non-personal data.
- 2. Causal analysis (Glymour & Herington, 2019; Salimi et al., 2019): Datasets used to train Al models often include relationships and dependencies between sensitive and non-sensitive variables. Thus, any attempts to mitigate bias in the dataset requires understanding the relationships between these variables. Otherwise, non-sensitive variables could act as proxies for the sensitive variables. Causal analysis helps with identifying these proxies, often in the form of visualizing as a graph the link between the variables in the dataset.

Causal analysis can be extended to "repair" the dataset by removing the dependencies based on pre-defined "fairness" criteria.¹ However, this approach relies on prior contextual knowledge about the AI model and its deployment, in addition to being computationally intensive for large datasets.

3. Transformation (Calmon et al., 2017; Feldman et al., 2015; Zemel et al., 2013): These approaches include transforming the data into a less biased representation. These transformations could involve editing the labels such that they become independent of specific protected groupings or based on specific "fairness" objectives.

Transformations are not without limitations. First, transformations usually affect the performance of the AI model and there is an inherent trade-off between bias mitigation and performance when using this approach. Second, transformations are limited to numerical data and cannot be used for other kinds of datasets. Third, this approach is susceptible to bias persisting due to the existence of proxy variables. For this reason, the use of this approach should be preceded by causal analysis to understand the links between the special category data and the proxy variables in the starting dataset. Even then, there is no guarantee that the transformations have eliminated the relationship between the special category data and

¹ The technical literature uses the term "fairness" and there are numerous definitions and metrics of "fairness" (Hutchinson & Mitchell, 2019). Many of these have been developed in the context of the USA, some based on the "four-fifths rule" from US Federal employment regulation, which are not valid in other contexts and countries (Watkins et al., 2022). Furthermore, these metrics are incompatible with each other (Kleinberg et al., 2016).

proxy variables. Third, transformations could make the AI model less interpretable (Lepri et al., 2018).

- 4. Massaging or relabeling (Kamiran & Calders, 2012): Relabeling is a specific type of transformation to strategically modify the labels in the training data such that the distribution of positive instances for all classes is equal. For example, if a dataset contains data about men and women, the proportion of the dataset that is labelled '+' for women should be the same as that for men. If the proportion is less for women, then some of the datapoints for women that were close to being classified as '+' but were initially labelled '-' will be changed, and the reverse will be done for datapoints for men. This approach is not restricted to training dataset and can also be used for validation and test datasets.
- Reweighing (Calders et al., 2009; Jiang & Nachum, 2020; Krasanakis et al., 2018): Instead of changing the labels in the dataset, this approach adds specific 'weight' for each data point to adjust for the bias in the training dataset. The weights can be chosen based on three factors:

 the special categories of personal data along with the probability in the population of this sensitive attribute, (2) the probability of a specific outcome [+/-] and (3) observed probability of this outcome for a sensitive attribute.

For instance, women constitute 50% of all humans, and if the label '+' is assigned to 60% of all data in the data set, then 30% of the dataset should contain women with a '+' label. However, if it is observed that only 20% of dataset has women with a '+' label, then a 1.5 weight is appended to women with a '+' label, 0.75 is appended to men with a '+' label, and so on, to adjust for the bias.

Alternatively, a more dynamic approach can be taken by training an unweighted classifier to learn the weights and then retrain the classifier by using those weights.²

Reweighing is more suitable for small models where retraining is not too expensive in terms of cost and resources.

- 6. Resampling (Kamiran & Calders, 2012): In contrast to the previous methods, the resampling method does not involve adding weights to the sample, nor does it involve changing labels in the training dataset. Instead, this approach focusses on how samples from the dataset are chosen to be used for training such that a balanced set of samples is used for training. Data from the minority class can be duplicated, or "oversampled", while data from the majority class can be skipped, or "under-sampled". The choice usually depends on the size of the entire dataset and the overall impact on the performance of the AI model. For instance, under-sampling requires datasets with sufficiently large amounts of data from the different classes.
- 7. Generating artificial training data (Sattigeri et al., 2019): When the quantity of available data is limited, especially for unstructured data such as images, a generative process can be used to develop the dataset. The use of generative adversarial networks (GAN) which includes specific bias considerations can contribute to generating and using less biased datasets for

² This process of training an unweighted model first, makes this approach of reweighing a mix of in-processing and pre-processing.

training. This approach assumes that an appropriate fairness criterion is available, which is a strong assumption, and it requires significant computing power.

1.2.2 In-processing

1. Regularisation (Kamishima et al., 2012): Regularisation is used in machine learning to penalise undesired characteristics. This approach was primarily used to reduce over-fitting but has been extended to address bias. This approach penalises classifiers with discriminatory behaviour. It is a data-driven approach that relies on balancing fairness (as defined by a chosen fairness metric) and a performance metric such as accuracy or the ratio between true positive rate and false positive rate for minority groups (Bechavod & Ligett, 2017).

While this approach is generic and flexible, it relies on the developer choosing the most suitable metric, which allows for gamification. In addition, there are also concerns that not all fairness measures are equally affected by regularisation parameters (Stefano et al., 2020). Furthermore, this approach could result in reduced accuracy and robustness.

2. Constrained optimisation (Agarwal et al., 2018; Zafar et al., 2017): Constrained optimisation, as the name suggests, constrains the optimisation function by incorporating a fairness metric during the model training by either adapting an existing learning paradigm or through wrapper methods. In essence, this approach changes the algorithm of the AI model. In addition to fairness metrics, other constraints that capture disparities in population frequencies can be included, resulting in trade-offs between the metrics.

The chosen fairness metric can result in vastly different models and hence, this approach is heavily reliant on the choice of the fairness metric, which results in difficulty to balance the constraints as well as unstable training.

3. Adversarial approach (Celis & Keswani, 2019; Zhang et al., 2018): While adversarial learning is primarily an approach to determine the robustness of machine learning models, it can also be used as a method to determine fairness. An adversary can attack the model to determine the protected attribute from the outputs. Then the adversary feedback can be used to penalise and update the model to prevent discriminatory outputs. The most common approach of incorporating this feedback is as an additional constraint in the optimisation process, that is, through constrained optimisation.

1.2.3 Post-processing

1. Calibration (Pleiss et al., 2017): Calibration is the process where the proportion of positive predictions is the same for all subgroups (protected or otherwise) in the data. This approach does not directly address the biases but tackles it indirectly by ensuring that the probability of positive outcomes is equal across social groups.

However, calibration is limited in flexibility and in accommodating multiple fairness criteria. In fact, the latter is shown to be impossible (Kleinberg et al., 2016). Although many approaches such as randomisation during post-processing have been suggested, this is an ongoing area of research without a clear consensus on the best approach.

2. Thresholding (Hardt et al., 2016; Kamiran et al., 2012): This approach recognises that most biases are close to the decision making boundary and that threshold rules rather than a hard cut-off can help reduce biased outcomes. One approach that has been suggested includes flipping the decisions within a certain threshold: giving favourable outcomes to unprivileged groups and unfavourable outcomes to privileged groups (Kamiran et al., 2012). Another approach, equalised odds, optimises for the ratio of true positive rate and false positive rate across all subgroups (Hardt et al., 2016). This approach is useful when historical bias is not present in the data. The threshold values themselves could be decided by a human or through statistical methods. The latter is a form of human-in-the-loop approach that allows for the human, who might be cognisant of the context of deployment to adjust the threshold values.

1.2.4 Methods for generative AI

The methods discussed so far have been developed for supervised learning, which relies on labelled data. Many of these methods can be used for facial recognition systems. However, the most recent generative AI models are developed self-supervised or unsupervised, without human labelling of the data used in the training process. While data provenance remains essential, additional methods have been suggested.

- 1. Data statements (Bender & Friedman, 2018): A data statement is a framework that goes a step further than data provenance to address bias issues when it comes to natural language processing and, by extension, to text generation AI. It includes information on annotator demography, source of data, languages it covers and the related demography, and the context of the data. A datasheet would include nuance such as whether a German text was collected from a website with high-German or Swiss-German.
- Fine-tuning the pre-trained model (Solaiman & Dennison, 2021): Some approaches to address bias in generative AI focus on fine-tuning the pre-trained model with desired characteristics. One approach includes the use of a carefully curated dataset that satisfies specific values (e.g., gender neutrality) to fine-tune the pre-trained model. Additional examples are added to this curated dataset based on the observed shortcomings in evaluations.
- 3. Modification to training (Keskar et al., 2019): Models can be trained such that the data are tagged to distinguish specific style, content or behaviour, which then results in outputs that satisfies these tags. This approach could potentially be used for training with tags that mitigate bias in the outputs.
- 4. Reinforcement learning with human feedback (RLHF) (Ziegler et al., 2020): After a model is pre-trained, the fine-tuning involves human annotators who rank (feedback) the generated output of this model. If certain kinds of gender biased outputs are encountered, the humans can give a low rank which would then be learnt by the model as something it should not output. In a way, this process can be thought of as an equivalent to labeling, but after the model has been pre-trained.

As mentioned earlier, RLHF has shortcomings and could itself introduce bias because it depends on the specific humans who are annotating, who may not be representative of the groups who are to be protected from bias (Casper et al., 2023).

Note that replacing gender specific words with gender neutral words (and their mapping) in word embedding is unlikely to mitigate bias (Gonen & Goldberg, 2019). Bias mitigation approaches for

generative AI is an ongoing research area and the methods listed in this section are proposals that are yet to be rigorously tested.

2 TOOLS FOR BIAS EVALUATION

Currently, no tool adequately detects and mitigates bias in generative AI systems. This is primarily because the state of the art is limited and is an ongoing research area. Thus, the tools below do not cater to generative AI, yet.

List of tools:

- 1. IBM AIF 360
- 2. Fairlearn
- 3. Holistic Al
- 4. Aequitas
- 5. What-If Tool
- 6. Other tools considered

2.1 IBM AIF360

IBM AIF360³ is limited in its scope but covers important bias detection and mitigation techniques pre-, in- and post-processing. However, it does not account for proxy variables, especially as one needs to specify the protected class.

Pre-processing techniques in the tool include reweighing and transformation, in-processing techniques include adversarial approach, and post-processing includes calibration and thresholding. Currently, the tool is limited in the number of important pre-processing methods.

This open-source tool is primarily designed for machine learning. It has been maintained for more than five years, has the possibility to update and add methods.

The use of this tool requires basic Python/R programming to make the best use. For instance, at least one small python code needs to be written per dataset whose bias is to be checked.

The tool can be run on a self-hosted instance and could potentially be useful for regulators. Companies, of course, can rely on their engineers to build on top of this tool.

2.2 Fairlearn

Fairlearn⁴ is a tool that was initially developed by Microsoft but has since been open-sourced and developed by a wider community. It is well documented and is the most thorough in explaining the

³ IBM AI Fairness 360 is part of a suite of tools developed and open-sourced by IBM. This tool focusses on bias and fairness. Other tools in this suite include AI Robustness 360 and AI Explainability 360. URL: https://github.com/Trusted-AI/AIF360/tree/master

⁴ URL: https://fairlearn.org/

logic behind the development of the tool as well as the limitations of the specific metrics and modules in the tool. It covers important bias detection and mitigation techniques—pre-, in- and postprocessing. In addition, it accounts for proxy variables through the thresholding mechanism.

Pre-processing techniques in the tool include transformation, in-processing techniques include adversarial approach, and post-processing includes thresholding. Currently, the tool is limited in the number of important pre-processing methods.

The use of this tool requires Python programming to make the best use. For instance, there will be a need to write python code to call the relevant parts of the software package and to make use of them. Examples in the form of Jupyter notebooks have been included to help with this.

The tool can be run on a self-hosted instance and could potentially be useful for regulators. This tool can be used to assess FRT. Only structured data for classification and regression is considered currently.

2.3 Holistic Al

HolisticAl⁵ provides AI governance products and services. For bias detection and mitigation, it also includes an open-source tool6, which is well documented and can be used as-is without the rest of the services offered by the company.

It covers important bias detection and mitigation techniques—pre-, in- and post-processing. Preprocessing techniques in the tool include reweighing and transformation, in-processing techniques include regularisation and constrained optimisation, and post-processing includes calibration and thresholding. Currently, this tool offers the largest set of state of the art techniques.

The use of the open-source tool requires Python programming. However, the company offers services that could make it possible to use the tool without programming experience.

It can be used for binary and multi-class classification regression and clustering.

2.4 Aequitas

Aequitas⁷ is an open-source tool designed to assist with detecting bias in AI systems. It is limited in its scope and has not seen much development in the past five years, primarily because it was the result of an academic project. However, the tool can be useful for developers of AI systems who can customise the tool for their purpose.

It covers pre-processing techniques such as transformation and massaging, and post-processing techniques such as thresholding. It also provides specific examples to detect bias in AI techniques such as logistic regression and random forest. However, it only covers binary classification.

2.5 What-If Tool

What-If Tool has been developed and maintained by Google. It can be used in several ways: through Jupyter notebooks, Tensorboard and on Google Cloud. Jupyter notebooks has the option of running the tool on the browser. Google claims that it does not "store, collect or share datasets" when this tool is used on the browser.

⁵ URL: https://www.holisticai.com/

⁶ URL: https://github.com/holistic-ai/holisticai

⁷ URL: https://github.com/dssg/aequitas

Overall, this tool can be useful to explore and understand the dataset. It also allows to explore alternatives (counterfactuals), something that is not present in any of the other tools currently. If the model in tensor format is available, then bias test can also be run on it.

Overall, after an initial learning curve, this tool is the most user friendly and could be used with minimal or no coding. However, the tool is not designed to cover a wide range of bias mitigation.

2.6 Other tools considered

- Amazon Sage Clarify: This tool has a range of bias detection and mitigation mechanisms. However, it can be used on Amazon AWS only.
- **Microsoft Responsible AI toolbox and AzureML:** Incorporates the open-source tool Fairlearn for bias detection and mitigation into a larger suite for AI governance. AzureML, as the name suggests, primarily caters to the users of Microsoft's cloud service—Azure.
- Secunet Antibias tool: This tool is in the development stage (as of December 2023) but intends to incorporate synthetic data into facial image dataset with the intention to improve performance of facial recognition system across groups. At this stage, two documents are available describing a proposed approach.⁸ However, no tool is available yet. As might be obvious, the potential future tool will be restricted to facial images and cannot be used for any other kinds of AI systems.

CONCLUSION

When personal data, including pseudonymised data, is used for the development of AI systems, data protection obligations apply. As a baseline, AI models should not use personal data such as first name, last name, date of birth, address and special categories of personal data, except when allowed under the EU AI Act for bias detection and correction. In addition, it is important to be cognisant of proxies that can allow for the inference of personal data and be the cause of bias.

Biased AI systems can harm people and mitigating bias is important. Understanding the sources of bias and data provenance is essential to mitigate bias. Various technical approaches at different stages of AI system development have been proposed to address bias. Open-source tools that include some of these approaches are available. These tools are at various stages of development and most of them require programming skills to use effectively. However, these technical approaches and tools should be complemented with contextual and socio-cultural knowledge as AI systems are not purely technical, but socio-technical.

⁸ URL: <u>https://arxiv.org/abs/2305.19962</u> and <u>https://arxiv.org/abs/2311.10476</u>

BIBLIOGRAPHY

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., & Wallach, H. (2018). A Reductions Approach to Fair Classification. *ICML*, *80*, 60–69.

Bagdasaryan, E., Poursaeed, O., & Shmatikov, V. (2019). Differential Privacy Has Disparate Impact on Model Accuracy. *NeurIPS*, 15453–15462.

Bechavod, Y., & Ligett, K. (2017). Penalizing unfairness in binary classification. *arXiv Preprint arXiv:1707.00044*.

Bender, E. M., & Friedman, B. (2018). Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics*, *6*, 587–604. https://doi.org/10.1162/tacl_a_00041

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *FAccT*, 610–623.

Bickel, P. J., Hammel, E. A., & O'Connell, J. W. (1975). Sex Bias in Graduate Admissions: Data from Berkeley: Measuring bias is harder than is usually assumed, and the evidence is sometimes contrary to expectation. *Science*, *187*(4175), 398–404.

Birhane, A., & Prabhu, V. U. (2021). Large image datasets: A pyrrhic win for computer vision? *WACV*, 1536–1546.

Blyth, C. R. (1972). On Simpson's paradox and the sure-thing principle. *Journal of the American Statistical Association*, *67*(338), 364–366.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Conference on Fairness, Accountability and Transparency*, 77–91.

Buyl, M., & De Bie, T. (2024). Inherent Limitations of AI Fairness. *Communications of the ACM*, *67*(2), 48–55. https://doi.org/10.1145/3624700

Calders, T., Kamiran, F., & Pechenizkiy, M. (2009). Building Classifiers with Independency Constraints. *ICDM Workshops*, 13–18.

Calmon, F. P., Wei, D., Vinzamuri, B., Ramamurthy, K. N., & Varshney, K. R. (2017). Optimized Pre-Processing for Discrimination Prevention. *NIPS*, 3992–4001.

Casper, S., Davies, X., Shi, C., Gilbert, T. K., Scheurer, J., Rando, J., Freedman, R., Korbak, T., Lindner, D., Freire, P., Wang, T., Marks, S., Segerie, C.-R., Carroll, M., Peng, A., Christoffersen, P., Damani, M., Slocum, S., Anwar, U., ... Hadfield-Menell, D. (2023). *Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback*. https://doi.org/10.48550/ARXIV.2307.15217

Celis, L. E., & Keswani, V. (2019). Improved Adversarial Learning for Fair Classification. *CoRR*, *abs/1901.10443*.

Cheney, J., Chiticariu, L., & Tan, W. C. (2009). Provenance in Databases: Why, How, and Where. *Found. Trends Databases*, 1(4), 379–474.

Crawford, K., & Paglen, T. (2021). Excavating AI: The politics of images in machine learning training sets. *AI & SOCIETY*. https://doi.org/10.1007/s00146-021-01162-8

d'Alessandro, B., O'Neil, C., & LaGatta, T. (2017). Conscientious Classification: A Data Scientist's Guide to Discrimination-Aware Classification. *Big Data*, *5*(2), 120–134.

Danks, D., & London, A. J. (2017). Algorithmic Bias in Autonomous Systems. *Ijcai*, *17*(2017), 4691–4697.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. S. (2012). Fairness through awareness. *ITCS*, 214–226.

European Union Agency for Fundamental Rights. (2022). *Bias in algorithms – Artificial intelligence and discrimination*. https://fra.europa.eu/en/publication/2022/bias-algorithm

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and Removing Disparate Impact. *KDD*, 259–268.

Ferry, J., Aïvodji, U., Gambs, S., Huguet, M.-J., & Siala, M. (2023). *SoK: Taming the Triangle -- On the Interplays between Fairness, Interpretability and Privacy in Machine Learning* (arXiv:2312.16191). arXiv. http://arxiv.org/abs/2312.16191

Fioretto, F., Tran, C., Hentenryck, P. V., & Zhu, K. (2022). Differential Privacy and Fairness in Decisions and Learning Tasks: A Survey. *IJCAI*, 5470–5477.

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proc. Natl. Acad. Sci. USA*, *115*(16), E3635–E3644.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H. M., III, H. D., & Crawford, K. (2018). Datasheets for Datasets. *CoRR*, *abs/1803.09010*.

Glymour, B., & Herington, J. (2019). Measuring the Biases that Matter: The Ethical and Casual Foundations for Measures of Fairness in Algorithms. *FAT*, 269–278.

Gonen, H., & Goldberg, Y. (2019). Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them. *NAACL-HLT (1)*, 609–614.

Gould, S. J. (1996). *Mismeasure of man*. WW Norton & company.

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, 29.

Hooker, S. (2021). Moving beyond "algorithmic bias is a data problem". *Patterns*, 2(4).

Hooker, S., Moorosi, N., Clark, G., Bengio, S., & Denton, E. (2020). *Characterising Bias in Compressed Models* (arXiv:2010.03058). arXiv. http://arxiv.org/abs/2010.03058

Hutchinson, B., & Mitchell, M. (2019). 50 Years of Test (Un)fairness: Lessons for Machine Learning. In danah boyd & J. H. Morgenstern (Eds.), *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019* (pp. 49–58). ACM. https://doi.org/10.1145/3287560.3287600

Jiang, H., & Nachum, O. (2020). Identifying and Correcting Label Bias in Machine Learning. *AISTATS*, *108*, 702–712.

Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, *33*(1), 1–33. https://doi.org/10.1007/s10115-011-0463-8

Kamiran, F., Karim, A., & Zhang, X. (2012). Decision theory for discrimination-aware classification. *2012 IEEE 12th International Conference on Data Mining*, 924–929.

Kamishima, T., Akaho, S., Asoh, H., & Sakuma, J. (2012). Fairness-Aware Classifier with Prejudice Remover Regularizer. *ECML/PKDD (2), 7524*, 35–50.

Keskar, N. S., McCann, B., Varshney, L. R., Xiong, C., & Socher, R. (2019). *CTRL: A Conditional Transformer Language Model for Controllable Generation* (arXiv:1909.05858). arXiv. http://arxiv.org/abs/1909.05858

Kleinberg, J. M., Mullainathan, S., & Raghavan, M. (2016). Inherent Trade-Offs in the Fair Determination of Risk Scores. *CoRR*, *abs/1609.05807*.

Krasanakis, E., Xioufis, E. S., Papadopoulos, S., & Kompatsiaris, Y. (2018). Adaptive Sensitive Reweighting to Mitigate Bias in Fairness-aware Classification. *WWW*, 853–862.

Kulynych, B., Yaghini, M., Cherubin, G., Veale, M., & Troncoso, C. (2022). Disparate Vulnerability to Membership Inference Attacks. *Proc. Priv. Enhancing Technol.*, 2022(1), 460–480.

Lepri, B., Oliver, N., Letouzé, E., Pentland, A., & Vinck, P. (2018). Fair, transparent, and accountable algorithmic decision-making processes: The premise, the proposed solutions, and the open challenges. *Philosophy & Technology*, *31*, 611–627.

Merler, M., Ratha, N., Feris, R. S., & Smith, J. R. (2019). *Diversity in Faces* (arXiv:1901.10436). arXiv. http://arxiv.org/abs/1901.10436

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447–453.

Paullada, A., Raji, I. D., Bender, E. M., Denton, E., & Hanna, A. (2021). Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns*, *2*(11), 100336. https://doi.org/10.1016/j.patter.2021.100336

Peng, K. (2020). Facial recognition datasets are being widely used despite being taken down due to ethical concerns. *Here's How*.

Plank, B., Hovy, D., & Søgaard, A. (2014). Linguistically debatable or just plain wrong? ACL (2), 507–511.

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J. M., & Weinberger, K. Q. (2017). On Fairness and Calibration. *NIPS*, 5680–5689.

Raji, I. D., & Buolamwini, J. (2019). Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 429–435.

Sagawa, S., Raghunathan, A., Koh, P. W., & Liang, P. (2020). An Investigation of Why Overparameterization Exacerbates Spurious Correlations. *ICML*, *119*, 8346–8356.

Salimi, B., Rodriguez, L., Howe, B., & Suciu, D. (2019). Interventional Fairness: Causal Database Repair for Algorithmic Fairness. *SIGMOD Conference*, 793–810.

Salon, O. (2019, March 12). *Facial recognition's 'dirty little secret': Millions of online photos scraped without consent*. NBC News. https://www.nbcnews.com/tech/internet/facial-recognition-s-dirty-little-secret-millions-online-photos-scraped-n981921

Sattigeri, P., Hoffman, S. C., Chenthamarakshan, V., & Varshney, K. R. (2019). Fairness GAN: Generating datasets with fairness properties using a generative adversarial network. *IBM J. Res. Dev.*, *63*(4/5), 3:1-3:9.

Schwartz, R., Vassilev, A., Greene, K., Perine, L., Burt, A., & Hall, P. (2022). *Towards a standard for identifying and managing bias in artificial intelligence* (NIST SP 1270; p. NIST SP 1270). National Institute of Standards and Technology (U.S.). https://doi.org/10.6028/NIST.SP.1270

Shankar, S., Halpern, Y., Breck, E., Atwood, J., Wilson, J., & Sculley, D. (2017). *No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World* (arXiv:1711.08536). arXiv. http://arxiv.org/abs/1711.08536

Solaiman, I., & Dennison, C. (2021). Process for Adapting Language Models to Society (PALMS) with Values-Targeted Datasets. *NeurIPS*, 5861–5873.

Stefano, P. G. D., Hickey, J. M., & Vasileiou, V. (2020). Counterfactual fairness: Removing direct effects through regularization. *CoRR*, *abs/2002.10774*.

Tan, Y. C., & Celis, L. E. (2019). Assessing Social and Intersectional Biases in Contextualized Word Representations. *NeurIPS*, 13209–13220.

Wachter, S., Mittelstadt, B., & Russell, C. (2020). Why Fairness Cannot Be Automated: Bridging the Gap Between EU Non-Discrimination Law and AI. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3547922

Watkins, E. A., McKenna, M., & Chen, J. (2022). The four-fifths rule is not disparate impact: A woeful tale of epistemic trespassing in algorithmic fairness. *CoRR*, *abs/2202.09519*. https://arxiv.org/abs/2202.09519

Weerts, H., Xenidis, R., Tarissan, F., Olsen, H. P., & Pechenizkiy, M. (2023). Algorithmic Unfairness through the Lens of EU Non-Discrimination Law: Or Why the Law is not a Decision Tree. *2023 ACM Conference on Fairness, Accountability, and Transparency,* 805–816. https://doi.org/10.1145/3593013.3594044

Zafar, M. B., Valera, I., Gomez-Rodriguez, M., & Gummadi, K. P. (2017). Fairness Constraints: Mechanisms for Fair Classification. *AISTATS*, *54*, 962–970.

Zemel, R. S., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013). Learning Fair Representations. *ICML* (3), 28, 325–333.

Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). Mitigating Unwanted Biases with Adversarial Learning. *AIES*, 335–340.

Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., Christiano, P., & Irving, G. (2020). *Fine-Tuning Language Models from Human Preferences* (arXiv:1909.08593). arXiv. http://arxiv.org/abs/1909.08593

