Machine Learning Bias in Computer Vision: Why do I have to care?

Camila Laranjeira
Computer Science Department
Universidade Federal de Minas Gerais
Belo Horizonte, Brazil
camilalaranjeira@dcc.ufmg.br

Virgínia Fernandes Mota
Colégio Técnico
Universidade Federal de Minas Gerais
Belo Horizonte, Brazil
virginia@teiacoltec.org

Jefersson Alex dos Santos
Computer Science Department
Universidade Federal de Minas Gerais
Belo Horizonte, Brazil
jefersson@dcc.ufmg.br

***Abstract*—Machine Learning bias is an issue with two main disadvantages. It compromises the quantitative performance of a system, and depending on the application, it may have a strong impact on society from an ethical viewpoint. In this work we inspect the literature on Computer Vision focusing on human-centered applications such as computer-aided diagnosis and face recognition to outline several forms of bias, bringing study cases for a more thorough inspection of how this issue takes form in the field of machine learning applied to images. We conclude with proposals from the literature on how to solve, or at least minimize, the impacts of bias.**

I. Introduction

Bias is one of the main issues faced by machine learning (ML) researchers. Practically speaking, models will always incorporate some form of bias, since it is unrealistic to build a complete representation of the real world. Thus, it is important to note that when we talk about bias, it refers to a lack of representation about one or more subsets of the real world, leading to skewed results, low-level quantitative performance, or analytical errors. If we, as machine learning researchers, want to solve real-world problems it is important to aim towards real-world representations. After all, we are all familiar with the "garbage in, garbage out" principle, in which poor data or algorithmic choices will most likely lead to poor results.

Being it an old Nikon camera that believes a Taiwanese-American woman is blinking[1], or the current British passport system saying an African-American woman has her mouth open[2] (refer to Fig. 1), machine-learning bias is currently leading to worldwide discussions on the importance of representation in artificial intelligence systems. That is an alarming red flag on our behavior as researchers and developers, since our actions can have a direct impact on society, hence we need to pay a lot more attention to the models we release.

Bias can occur in a variety of stages of the machine learning pipeline, from human reporting and selection bias to algorithmic and interpretation bias, as we will cover more thoroughly throughout this work. Therefore, the solution is not simple and requires analyzing the bias origin. Therefore, the objective of this work is to discuss bias in machine learning models, how it affects the inferences, and how we can solve or at least minimize its impacts. Mainly because bias not only affects the performance of a model based on quantitative measures such as accuracy, but it can also extend to issues of ethics, fairness, and inclusion.

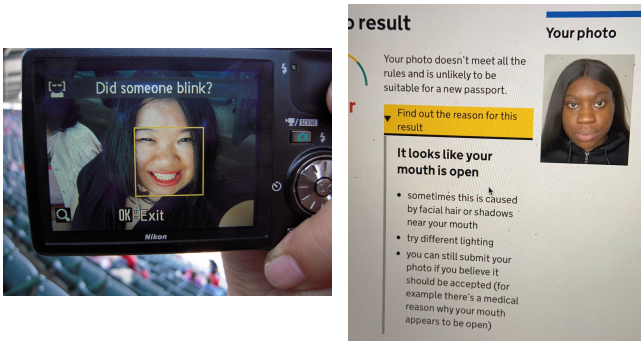

Fig. 1. Examples of social impacts caused by biased models. Photos extracted from The Society Pages[1] and Twitter[2].

The rest of the paper is organized as follows: Section II describes the concept of bias, types of worrisome bias in computer vision (CV), and draws insights from the CV literature, highlighting potential sources of bias; Section III presents recent works focusing in mitigating ML bias; Section IV summarizes where this research field is moving towards; and finally, Section IV presents the conclusion.

II. Machine Learning Bias: Why do I have to care?

This section describes the concept of bias, the most common and most worrisome types of bias in computer vision, and draws insights from the CV literature, highlighting potential sources of bias.

A. What is Bias?

Bias, in our context, must not to be confused with the bias term in machine learning models or prediction bias. Regarding the bias term, its mathematical definition is an intercept or offset from an origin. Bias is referred to as b or w_0 in machine learning models.

For example, bias is the b in the following formula:

$$f(x) = b + w_1x_1 + w_2x_2 + ... + w_nx_n \quad (1)$$

[1]https://thesocietypages.org/socimages/2009/05/29/nikon-camera-says-asians-are-always-blinking/

[2]https://twitter.com/elainebabey/status/1232333491607625728

As for prediction bias, it is a quantity that measures how far apart the prediction mean is from the real world mean. Large biases in this case can be caused by noisy data, faulty pipeline, or even biased training samples, to name a few. Thus, machine learning bias, from an ethical point of view, could cause prediction bias. We will see more in the next section.

Fairness in machine learning refers to the attempt to correct algorithm bias. That is, for fairness, the answers a machine learning algorithm gives must not fall foul of bias or discrimination. Therefore, what we call Machine Learning Bias in this paper is the term used in ethics/fairness for:

1. Stereotyping, prejudice or favoritism towards some things, people, or groups over others. These biases can affect collection and interpretation of data, the design of a system, and how users interact with a system. 2. Systematic error introduced by a sampling or reporting procedure.: (Merriam-Webster Dictionary definition).

In other words, it is a type of error where certain sets of the real world are more weighted and/or represented than others. That type of error can lead to Algorithmic Bias, that is, any kind of Algorithmic Solution that is unjust, unfair or prejudicial [1]. As machine learning becomes more and more pedestrian and common in society daily activities, we need to remember that an algorithm is written by humans using data collected for human purposes.

On a machine learning straightforward pipeline, we often see the following steps: data is collected, annotated and goes through preprocessing; model is trained; results are aggregated and analysed according to the task. Then, where can the bias be along that pipeline?

Even before collecting the data for the first step of that pipeline, the data itself can host a lot of human biases, such as stereotyping, prejudice or racism. Then, as we collect and annotate this data, we can introduce more biases, such as sampling errors, in-group and out-group bias, and so on. The algorithmic choices we make regarding the model, for instance loss functions or regularization terms, can exacerbate any preexisting inclinations on the data. And finally, how we analyse and present the results can also lead to bias in interpretation.

From social inclinations present on the available data, to our own personal biases through the process of building and releasing a model, every step of a machine learning pipeline is subject to introducing some form of bias.

B. Types of Bias

In this section we will describe the most common, and most worrisome, forms of bias in computer vision. It is important to note that there are other forms of bias as well as subtypes of the ones listed below.

Selection bias. It occurs when data collection is not properly randomized, leading to a lack of diversity in the training population with respect to the real world scenario. There are several types of selection bias, such as cherry picking (i.e. intentionally choosing which data to collect), incorrect partitioning of data, or most commonly sample bias. Recent works in the area of face recognition highlights **sample bias**, with databases overwhelmingly skewed towards white people's pictures, leading to lower performance on people of different ethnicities [2], [3]. It is important to note that any system meant to work with or for people, should account to a more complete representation of the chosen population. However the large availability of public images portraying white people, usually male, along with search engines biased towards the same population [4] leads to a vicious cycle of replicating sample bias.

Automation bias. It refers to an overdependence on automated systems, to a degree that correct decisions are overlooked in favor of automated ones. It not only refers to society as it increasingly relies on technology, but also to the developers and researchers blindly automating the development of such technology. A recent study [5] analysed one of the most important datasets in the field of computer vision, ImageNet. In the age of Deep Learning, researchers invest a lot of time on the collection of large datasets. But that can only go so far, thus major steps in the process of data collection rely on automation. The authors showed how it led ImageNet to replicate harmful biases, such as misogynistic labels for women's photos, as well as included clearly non-consensual images (e.g. upskirt).

Measurement bias. As the name implies, it refers to faulty, low quality or unreliable measures when collecting data. For example, it can come from equipment choices, such as collecting images with different cameras. And it is very salient in the field of computer vision in the form of inconsistent or unreliable labeling of samples, which can have many causes such as insufficient label options (e.g. binary gender [6]), labeling of abstract concepts (e.g. sentiments [7]), cultural biases [8], or even poor training of those responsible for the labeling. Whenever label inconsistencies arise from subjective views from labelers, it is related to the concept of **recall bias**, a subtype of measurement bias.

Expectation bias. It is important to highlight that unconscious expectations or preexisting worldviews can, and probably will, impact our choices and conclusions throughout the process of research. The work of [9] is a very interesting reading on how records of expectation bias dates back to Newton's theories, along with multiple examples of how careless one can get once the results meets their expectations. Expectation bias may interfere with experiments through **confirmation bias**, when researchers favor data or results in accordance with their beliefs.

Evaluation bias. A definition of evaluation bias is given in [10], relating it both to misrepresentative benchmarks as well as poor metrics widely accepted in a given literature. Often, the choice of metrics does not account for diversity aspects, for instance aggregated means can hide underperformance in certain groups of the population. Once benchmarks and metrics are established, the following researchers on the same topic will tend to adopt them since they need to compare their work with the present literature. Evaluation bias was mentioned by authors of [11] in a bias section later added to the original

paper, after their work was subjected to public criticism on its racial bias.

Overgeneralization. The assumption that a certain analysis, knowledge, or in our case, machine learning model, will apply to broader scenarios. It relates to the concept of overfitting, in the sense that models fit on a certain subset might not generalize well to different scenarios. We could refer once again to the example of biased face recognition systems, since authors usually assert that the model is supposed to recognize people as a general population, when they should specify the ethnicity majority the model was trained on. Note that the sample bias, where we first mentioned face recognition, happens at data collection stages, while overgeneralization refers to conclusions drawn from the study. Another example worth mentioning is a work from the field of social psychology, which concludes that it is possible to estimate sexual orientation in the general population based on the high accuracy achieved in a specific, and highly skewed, dataset [12].

Racial bias. It can occur on several stages of the research process, from data collection to conclusions drawn by human analysts. Racial bias can take many forms, for instance the sample bias in computer aided systems to detect skin cancer lesions, in which the majority of lesions are from white skinned individuals [13], or the automation bias when a risk assessment system was deployed in a criminal recidivism evaluation [14], with human analysts disregarding the evident bias towards criminalizing black individuals. The work of [10] expands on the concept of racial bias, outlining the **Historical bias**, in which many forms of stereotypes are contemplated.

C. Study Cases

In this section we draw insights from the literature of computer vision, highlighting potential sources of bias.

Detecting and recognizing different face attributes has become an increasingly feasible machine learning task due to the rise of deep learning techniques and large people-focused datasets [3]. One of the most popular faces dataset is CelebA. CelebFaces Attributes Dataset (CelebA) [2] is a large-scale face attributes dataset with more than 200K celebrity images, each with 40 attribute annotations. The images in this dataset cover large pose variations and background clutter. Authors claim CelebA has large diversity, large quantity, and rich annotations.

However, the work of Buolamwini and Gebru [15] puts a spotlight on a major issue: commercial face recognition models have a considerable decrease in performance for people with darker skins, specially the ones labelled as female. Indeed, several commercial and state-of-the-art models are trained in datasets such as CelebA, which lack the diversity to represent real world people.

One example is PULSE [11]. PULSE (Photo Upsampling via Latent Space Exploration), is an algorithm which generates high-resolution, realistic images at resolutions previously unseen in the literature. It was trained in CelebA HQ (a high-quality version of CelebA that consists of 30,000 images at 1024×1024 resolution). As depicted in Figure 3, it can create

Fig. 2. CelebFaces Attributes Dataset (CelebA) examples of images and annotations [2].

Fig. 3. PULSE results example: (x32) The input (top) gets upsampled to the SR image (middle) which downscales (bottom) to the original image [11].

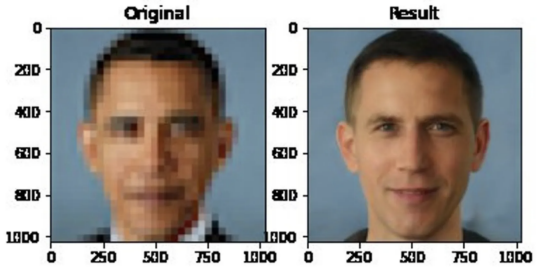

Fig. 4. PULSE results on Barack Obama's face. Picture extracted from Twitter[3].

realistic images from poor quality images. However, in reply to the author's Twitter post, many people highlighted the poor performance of the model on dark skin individuals, as depicted in Figure 4, extracted from one Twitter reply[3]. In a model card later added to the paper by the authors, they attribute the poor performance to the training/validation data, claiming that it is the accepted benchmark in the field, leading to an evaluation bias.

We can relate both cases, face recognition and photo upsampling, with selection bias, evaluation bias and racial bias. Unfortunately, those are not uncommon to find in the literature, as well as commercial systems. Google's image recognition tools have returned racially biased results since 2015[4]. In that year, Google Photos labelled two dark-skinned individuals as "gorillas". Although the company released a public apology, according to a report by Wired[5], Google did not fix the issue. Instead, it simply stopped returning the "gorilla" label, even for pictures of that specific mammal. They later banned the words "chimp," "chimpanzee," and "monkey". In 2020, Google Cloud Vision still has major issues with racial bias, as another report by AlgorithmWatch[6] found. As Fig. 5 shows, the tool labeled an image of a dark-skinned individual holding a thermometer as "gun", while the same image with a simple editing strategy to lighten the skin was labeled as "monocular". After the article was published, the service stopped returning the label "gun", although we can not say for sure what type of measures were taken, since there is little transparency when it comes to commercial systems. In 2021, AlgorithmWatch is still reporting issues with mislabeling dark-skinned individuals as animals, even in cartoons[7]. This time Apple, as well as Google, were caught making the same mistake.

Another application that analyses facial attributes and relates to racial bias is: Predicting criminal recidivism. In 2016, Julia Angwin, a reporter from ProPublica, wrote on how machine learning software used across the United States to predict criminal recidivism are biased towards black people[8]. The system is called Correctional Offender Management Profiling for Alternative Sanctions (COMPAS), and it incorporates a range of supposedly relevant criminogenic factors emerging from meta-analytic studies of recidivism. In this case, the data itself is inherently biased, as racism is a major issue in the United States' criminal system. It strongly relates to the definition of historical bias given in [10], since the data may accurately represents the world, but still reflect historical prejudices towards a specific group. Additionally, measurement bias plays an important role, since the chosen criminogenic factors are proxies for socioeconomic status, such as ZIP code, household income, etc.

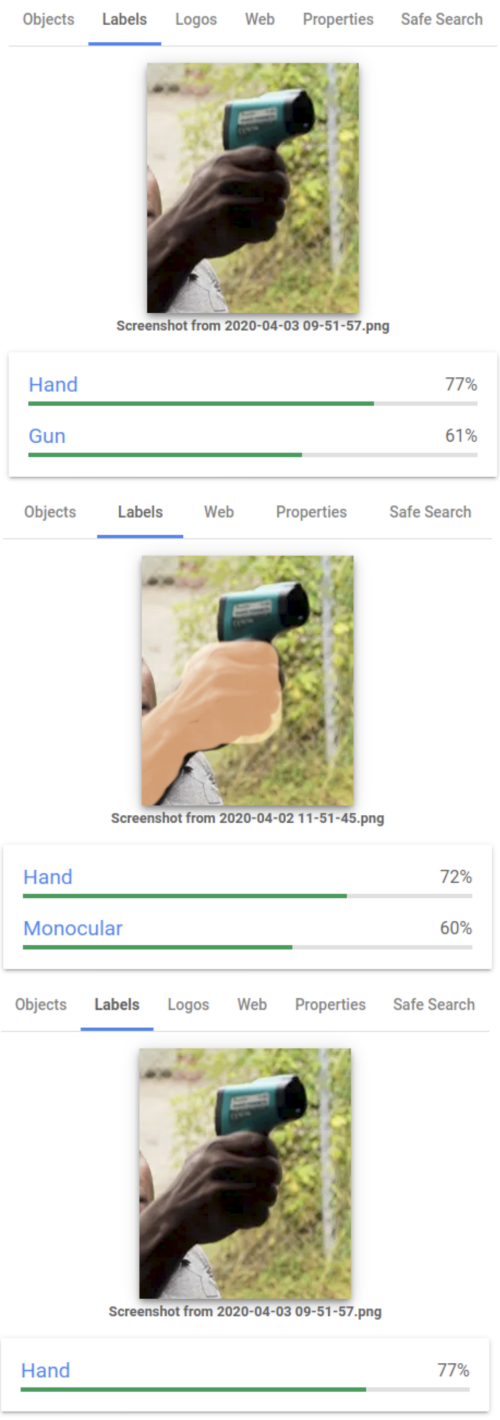

Fig. 5. Google Vision Cloud results for an hand-held thermometer on a black hand (first image) and on a white hand (second image). After the publication of the AlgorithmWatch article, the service does not return a "gun" label anymore (third image). Retrieved from AlgorithmWatch[6].

[3]https://twitter.com/Chicken3gg/status/1274314622447820801?s=20
[4]https://www.bbc.com/news/technology-33347866
[5]https://www.wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind/
[6]https://algorithmwatch.org/en/google-vision-racism/
[7]https://algorithmwatch.org/en/apple-google-computer-vision-racist/
[8]https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

In order to analyse the effectiveness of COMPAS, ProPublica obtained the risk scores assigned to more than 7,000 people arrested in Florida and checked how many were charged with new crimes over the following two years. They found that only 20% of people scored as potentially dangerous actually relapsed into criminal activities. Moreover, they found that black defendants were still 77% more likely to be pegged as higher risk of committing a future violent crime, 45% more likely for crimes of any kind, while white defendants were mislabeled as low risk more often than black defendants. We can relate the usage of such systems to automation bias, since users blindly relied on outcomes from the system. It might even relate to confirmation bias, if users were favoring automated decisions due to their agreement with such decisions.

Racial biases can also be found in medical image databases, as authors in [13] evaluated in the context of skin lesions detection. Although skin cancer is more common for white individuals, evidence show that diagnosis for people of color may only occur in advanced stages of the disease, leading to a lower survival rate within the population [16]. As authors in [13] showed, benchmark datasets are overwhelmingly biased towards "light" categories of the Fitzpatrick skin phototype classification, which may lead researchers to evaluation bias.

A highly unethical study was conducted by Wang and Kosinski [12], where they claim to have created a "sexual orientation detector" using $35,326$ images from public profiles on a US dating website. Authors alleged that their findings were "Consistent with the Prenatal Hormone Theory (PHT) of sexual orientation", later concluding that "gay men and women tend to have gender-atypical facial morphology". That is both a spurious correlation and an overgeneralization, since the study does not cover causality in the relationship between sexual orientation and facial attributes, and the authors can not claim that their model is able to rate sexual orientation outside the population from the study.

The so called "sexual orientation detector" also shows clear signs of selection bias. As depicted in Figure 6, the "average" straight woman appears to wear eyeshadow, while the "average" lesbian does not. Glasses are visible on the gay man, and to a lesser extent on the lesbian, while they seem absent in the heterosexual composites. It seems the algorithm's ability to detect orientation has little to do with facial structure, but actually refers to grooming, presentation and lifestyle.

Among many tools for automatic data collection, image search engines are often used in the field of Computer Vision. As research suggests, there are many issues with the ranking system responsible for outputting results in a certain order from a query [17], leading to the propagation of multiple stereotypes. Findings in [5] show severe problems with ImageNet, an automated collection from search engines that is widely used as a pretraining database. Gender stereotypes are clearly presented on the database, with misogynistic labels associated to female presented individuals due to automation bias. Researchers have been studying search engine algorithms as the source of such gender stereotypes in databases, finding not only stereotype exaggerations by professional career [18], as well a skewed gender distribution for specific trait adjectives in a paper cleverly entitled "Competent Men and Warm Women" [19].

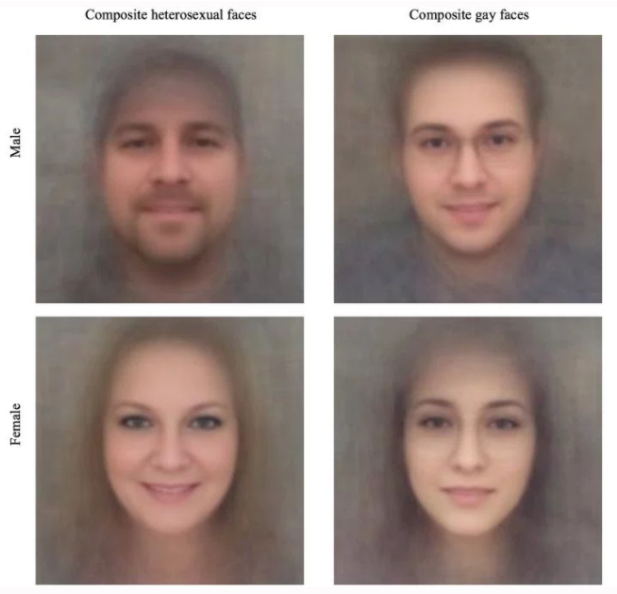

Fig. 6. Average faces from the dataset collected by the authors of [12].

III. Debiasing: How to solve ML Bias?

Regarding a straightforward machine learning solution, we saw that bias can be introduced in every step. Indeed, the prevention of bias is an ongoing process. Though it is sometimes difficult to know when your data or model is biased, there are several steps you can take to help prevent bias or catch it in the early stages of a project.

One of the main points to prevent data bias is to ensure that your research team, data scientists and labelers are diverse. Likewise, wherever possible, combine inputs from multiple sources to ensure data diversity. These actions could mitigate any human biases in data annotation and collection. Additionally, ask the help of someone with domain expertise to review your collected and/or annotated data. Someone outside your team may see biases that your team has overlooked. Real-world problems are often multidisciplinary [20].

To prevent measurement and recall bias, define clear guidelines for data labeling expectations. Moreover, create a gold standard for your data labeling. A gold standard is a set of data that reflects the ideal labeled data for your task. In that way, data labelers tend to be more consistent.

The work of [21] suggests using a multi-pass annotation system. That is, multiple humans (or models) place a label independently of one another, and only unanimous labels are assigned to the sample. Examples of this include sentiment analysis, content moderation and intention recognition. A dataset for computer-aided diagnosis of lung lesions, called CheXpert [22], is an example where validation samples are the very few in which a consensus was achieved by three

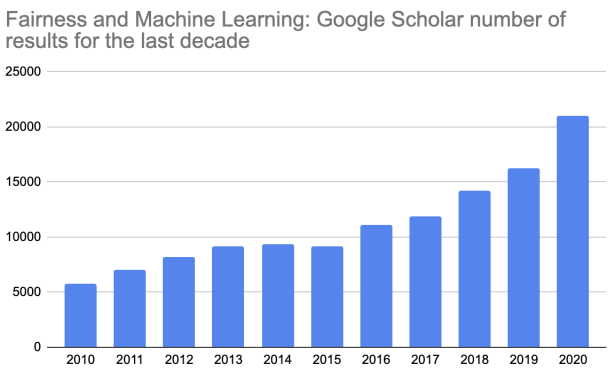

Fig. 7. Fairness and Machine Learning: Google Scholar number of results for the last decade.

radiologists. Training samples are assigned uncertainty labels instead.

Prevent and/or mitigate bias is an ongoing process, so analyze your data regularly. Keep track of errors and problem areas so you can respond to and resolve them quickly.

Therefore, when creating a dataset we need to understand our data and make this knowledge available to works that will use it. Since there is no true unbiased dataset, the least we can do is to learn what kind of bias we have to handle.

One approach for this is described in [23]. Authors propose datasheets for datasets, which may increase transparency and accountability within the machine learning community and mitigate unwanted biases in machine learning systems. The proposed workflow has seven main topics with questions to be answered by dataset creators in order to facilitate reproducibility of ML results, and help researchers select more appropriate datasets for their chosen tasks. The topics are: Motivation/Purpose of the dataset; Composition; Collection Process; Preprocessing/cleaning/labeling; Uses; Distribution; Maintenance.

It is interesting to note that datasheets can help not only mitigating data bias, but also to understand ethical and legal implications of the dataset, mainly dealing with those that will be publicly available. The mere action of writing a datasheet can instigate the authors to reflect on their work from a fairness perspective.

An interesting survey on what is fair and unfair and how it is applied to ML Fairness is depicted in [1]. Authors studied the 50-year history of fairness definitions in the areas of education and machine-learning. In [24] authors describe the many types of biases that occur in data and present the different ways that the concept of fairness has been studied and applied in literature. It is interesting to note that ML Fairness is a recent field and the number of publications has doubled since 2016 (Figure 7).

Since many computer vision models that use deep learning rely on transfer learning on pre-trained models, mitigating data bias on the source is not always practicable. There are some techniques to reduce the impact of such biases and to analyze how biased your data or your model is.

In our first steps into machine learning we learn that we should separate the data into train, validation, and test splits to prevent the model from overfitting and to accurately evaluate it. This model is then evaluated using one or more aggregate performance metrics, such as accuracy, precision, recall. However, those metrics can obscure poor performance for groups of people that are not well represented in an evaluation dataset. To evaluate fairness in these models, we can also apply disaggregated/intersectional evaluation [25], [26].

Disaggregated evaluation refers to a comparison across subgroups, evaluating the selected metrics on each of those groups in the population. Intersectional evaluation is a type of disaggregated evaluation which combines more than one subgroup and compares them across subgroups. If the recall is equal across the subgroups, we have equality in the dataset. Similarly, having the same prediction across subgroups is equivalent to Predictive Parity fairness criteria [27].

Let's use face detection as an example. Considering binary gender labels: female and male. We can look at each pair (female, face detection), (male, face detection); then evaluate how the error rates are different or similar. We can also look at this problem by analysing it intersectionally, creating tuples such as (black women, face detection), (white men, face detection). In fact, this is the basis of the study Gender Shades from [15]. Thus, we can see how well the system is doing across different types of individuals in the dataset. This analysis between groups can also be applied using alternative metrics, such as the disparate impact [28], which uses the accuracy proportion between unprivileged and privileged groups to fairly evaluate the model, instead of the classic measures such as accuracy, precision, recall, among others.

Other strategies to evaluate such proportion between groups are Demographic Parity (also called Independence, Statistical Parity) and Equality of Odds (also called Separation, Positive Rate Parity). Demographic Parity states that the proportion of each segment of a protected/sensitive class (e.g. gender, race) should receive the positive outcome at equal rates. A classifier satisfies the Equality of Odds criteria if the subjects in the protected and unprotected groups have equal true positive rate and equal false positive rate [29].

Choose your evaluation metrics considering acceptable trade-offs between False Positives and False Negatives. Each task will have a different fairness criteria to be applied [25]. An interesting comparison between fairness metrics is described in [30]. Authors claim that this metric selection can be guided by the observability of each statistic in practice. Indeed, authors in [31] highlight some important aspects about the relationships between fairness metrics, in particular with respect to the distinctions individual vs. group and observational vs. causality-based.

One example of a debiasing solution is applied for skin lesion datasets in [32]. Authors propose destructive and constructive actions in the target datasets, and raised the question "If we hide the lesion information from the network, can it still learn patterns that help differentiate benign from malignant lesions?". Surprisingly, even removing 70% of all pixels in

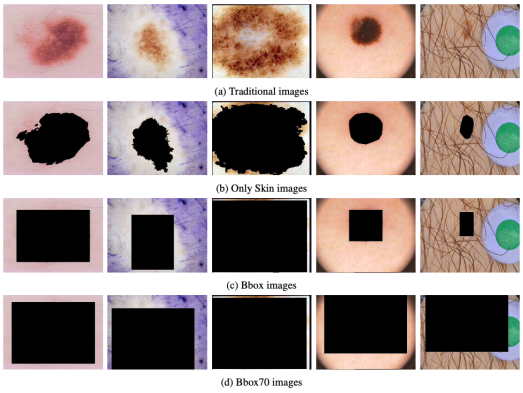

Fig. 8. Samples from each of disrupted datasets in [32].

the image and all medical relevant features that could aid the classification, the network was able to make decisions that are much better than chance (Figure 8). Therefore, the outcome results strongly relies on patterns introduced during image acquisition and general dataset bias. In the subsequent work, authors found out that, despite interesting results that point to promising future research, current debiasing methods are not ready to solve the bias issue for skin-lesion models [33].

After all this work to debias the data and the model, we can also focus on releasing the ML model in a responsible way. While datasheets for datasets [23] refers to the data itself, [34] describes Model Cards, short documents accompanying trained machine learning models that provide benchmark evaluation in a variety of conditions, such as across different cultural, demographic, or phenotypic groups (e.g., race, geographic location, sex, Fitzpatrick skin type) and intersectional groups (e.g., age and race, or sex and Fitzpatrick skin type) that are relevant to the intended application domains.

Big tech enterprises that are known for the use of Big Data are also creating tools to ensure fairness in ML. The AI Fairness IBM 360[9] is an open source toolkit available in Python and R to help examine, report, and mitigate discrimination and bias in machine learning models throughout the artificial intelligence application lifecycle [35]. In order to analyze your data in different situations, Google launched the What-If Tool (WIT)[10]. It is possible to test different scenarios and visualize model behavior across multiple models and subsets of input data, and for different ML fairness metrics.

Lastly, we should also be concerned with biases that might emerge in the deployment step. In this case, automation bias and confirmation bias are some of the major threats. Meredith Broussard mentions in her book, Artificial Unintelligence [36], the concept of "technochauvinism" referring to the excessive trust we place on technology to solve all problems in the world, while Cathy O'Neil's book [37] popularized important examples of injustices caused by excessive automation. It should be noted that if a researcher intends to develop solutions for human-centered applications, the role of experts in humanities is essential to craft solutions that account for the complexity of society and its individuals.

IV. Discussion and Future Works

There is no silver bullet for debias the data and the model or the interpretation of a result. Moreover, when talking about real world data we always bring real world bias. Even so, we need to pay attention that machine learning can unintentionally lead to unfair outcomes. The unfairness can come from different sources, thus we have to be aware of bias in our data, in our model, and in interpreting the results.

We have to have fairness, accountability, transparency and responsibility throughout the Machine Learning pipeline. The care with the dataset does not end when you launch it or publish it.

Works as [23], [34] are showing that it is possible to create a ML pipeline according to fairness and responsibility criteria. We can also expanding this idea to all ethical and legal implications that are involved with human-centered applications. In [38][11], authors found worrisome issues on several well-known used datasets. They found that ethical problems underlying a dataset can permeate into an ecosystem of derivatives, amplifying their effect and making it challenging to make effective corrections. Moreover, they found that modifications of the dataset through derivatives can introduce new harms.

In that sense, fairness ensures that the output of the machine does not have an unjust impact on end-users from any demographic. Accountability and Responsibility state that there's someone responsible for the results of AI-fuelled decisions. It's about being able to explain and control the outcome of such decisions. And, where harm occurs, that someone could be legally responsible. And finally, Transparency is about the ability to see and explain two key things: exactly what the machine learning algorithm has learned and how it uses what it's learned to reach its final output. This is also known as explainable artificial intelligence.

It is not enough to put human-centered technology as a vague overall score associated to it, we need to understand across different populations, how our model behaves, what the data is telling us and what ethical and legal implications we have to toil.

V. Conclusion

In summary, why do I have to care? Practically speaking, models will always incorporate some form of bias, since it is unrealistic to build a complete representation of the real world. Therefore, it is important to be aware of the potential biases in machine learning for any data project and how it can lead to unfair outcomes. In the real world, people are not just numbers to be evaluated, our decisions directly impact society and its individuals.

With this paper we hope to communicate with the computer vision community, raising awareness of the risks that machine learning bias poses to society. From data collection, annotation, and treatment, to every algorithmic choice we make,

[9]https://aif360.mybluemix.net/

[10]https://pair-code.github.io/what-if-tool/

[11]Also available in https://www.youtube.com/watch?v=1BAJMUNf1tM

it is our role to minimize the damages and maximize the benefits of technology. It is a major research challenge we hope the community will be interested in tackling. The future of machine learning for human-centered applications relies on fairness, accountability, transparency, and responsibility.

ACKNOWLEDGMENTS

This work was supported by the Serrapilheira Institute (grant number Serra – R-2011-37776). This work was supported in part by the Minas Gerais Research Funding Foundation (FAPEMIG) under Grant APQ-00449-17, by the National Council for Scientific and Technological Development (CNPq) under Grant 311395/2018-0 and Grant 424700/2018-2, and by the *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES)* – Finance Code 001.

REFERENCES

[1] B. Hutchinson and M. Mitchell, "50 years of test (un)fairness: Lessons for machine learning," in *FAT* '19: Conference on Fairness, Accountability, and Transparency*, 2019.

[2] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[3] H. J. Ryu, M. Mitchell, and H. Adam, "Inclusivefacenet: Improving face attribute detection with race and gender diversity," in *Workshop on Fairness, Accountability, and Transparency in Machine Learning (FAT/ML)*, 2018.

[4] A. Singh and T. Joachims, "Fairness of exposure in rankings," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018, pp. 2219–2228.

[5] V. U. Prabhu and A. Birhane, "Large image datasets: A pyrrhic win for computer vision?" *arXiv preprint arXiv:2006.16923*, 2020.

[6] M. K. Scheuerman, J. M. Paul, and J. R. Brubaker, "How computers see gender: An evaluation of gender classification in commercial facial analysis services," *Proceedings of the ACM on Human-Computer Interaction*, vol. 3, no. CSCW, pp. 1–33, 2019.

[7] J. Zeng, S. Shan, and X. Chen, "Facial expression recognition with inconsistently annotated datasets," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 222–237.

[8] C. Schumann, S. Ricco, U. Prabhu, V. Ferrari, and C. Pantofaru, "A step toward more inclusive people annotations for fairness," *arXiv preprint arXiv:2105.02317*, 2021.

[9] M. Jeng, "A selected history of expectation bias in physics," *American Journal of Physics*, vol. 74, no. 7, pp. 578–583, 2006.

[10] H. Suresh and J. V. Guttag, "A framework for understanding sources of harm throughout the machine learning life cycle," *arXiv preprint arXiv:1901.10002*, 2019.

[11] S. Menon, A. Damian, S. Hu, N. Ravi, and C. Rudin, "PULSE: self-supervised photo upsampling via latent space exploration of generative models," in *Computer Vision and Pattern Recognition (CVPR)*, 2020.

[12] Y. Wang and M. Kosinski, "Deep neural networks are more accurate than humans at detecting sexual orientation from facial images," *Journal of Personality and Social Psychology*, vol. 114(2), p. 246–257, 2018.

[13] N. M. Kinyanjui, T. Odonga, C. Cintas, N. C. Codella, R. Panda, P. Sattigeri, and K. R. Varshney, "Fairness of classifiers across skin tones in dermatology," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 320–329.

[14] W. Dieterich, C. Mendoza, and T. Brennan, "Compas risk scales: Demonstrating accuracy equity and predictive parity," *Northpointe Inc*, 2016.

[15] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, ser. Proceedings of Machine Learning Research, S. A. Friedler and C. Wilson, Eds., vol. 81. New York, NY, USA: PMLR, 23–24 Feb 2018, pp. 77–91. [Online]. Available: http://proceedings.mlr.press/v81/buolamwini18a.html

[16] M. Gohara, "Skin cancer: an african perspective," *British Journal of Dermatology*, vol. 173, pp. 17–21, 2015.

[17] R. Gao and C. Shah, "Toward creating a fairer ranking in search engine results," *Information Processing & Management*, vol. 57, no. 1, p. 102138, 2020.

[18] M. Kay, C. Matuszek, and S. A. Munson, "Unequal representation and gender stereotypes in image search results for occupations," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 2015, pp. 3819–3828.

[19] J. Otterbacher, J. Bates, and P. Clough, "Competent men and warm women: Gender stereotypes and backlash in image search results," in *Proceedings of the 2017 chi conference on human factors in computing systems*, 2017, pp. 6620–6631.

[20] B. Cowgill, F. Dell'Acqua, S. Deng, D. Hsu, N. Verma, and A. Chaintreau, "Biased programmers? or biased data? a field experiment in operationalizing ai ethics," in *ACM Conference on Economics and Computation*, 2020.

[21] D. M. Iraola and A. J. Yepes, "Single versus multiple annotation for named entity recognition of mutations," 2021.

[22] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya *et al.*, "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 590–597.

[23] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. M. Wallach, H. D. III, and K. Crawford, "Datasheets for datasets," *CoRR*, vol. abs/1803.09010, 2018. [Online]. Available: http://arxiv.org/abs/1803.09010

[24] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *CoRR*, vol. abs/1908.09635, 2019. [Online]. Available: http://arxiv.org/abs/1908.09635

[25] S. Barocas, M. Hardt, and A. Narayanan, *Fairness and Machine Learning*. fairmlbook.org, 2019.

[26] S. Barocas, A. Guo, E. Kamar, J. Krones, M. R. Morris, J. W. Vaughan, D. Wadsworth, and H. M. Wallach, "Designing disaggregated evaluations of AI systems: Choices, considerations, and tradeoffs," *CoRR*, vol. abs/2103.06076, 2021. [Online]. Available: https://arxiv.org/abs/2103.06076

[27] S. Verma and J. Rubin, "Fairness definitions explained," *ACM/IEEE International Workshop on Software Fairness*, 2018.

[28] S. Caton and C. Haas, "Fairness in machine learning: A survey," 2020.

[29] M. Du, F. Yang, N. Zou, and X. Hu, "Fairness in deep learning: A computational perspective," 2020.

[30] P. Garg, J. Villasenor, and V. Foggo, "Fairness metrics: A comparative analysis," 2020.

[31] A. Castelnovo, R. Crupi, G. Greco, and D. Regoli, "The zoo of fairness metrics in machine learning," 2021.

[32] A. Bissoto, M. Fornaciali, E. Valle, and S. Avila, "(de)constructing bias on skin lesion datasets," *CoRR*, vol. abs/1904.08818, 2019. [Online]. Available: http://arxiv.org/abs/1904.08818

[33] A. Bissoto, E. Valle, and S. Avila, "Debiasing skin lesion datasets and models? not so fast," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020.

[34] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Rajia, and T. Gebru, "Model cards for model reporting," in *FAT* '19: Conference on Fairness, Accountability, and Transparency*, 2019, pp. 220–229.

[35] R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic, S. Nagar, K. N. Ramamurthy, J. T. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, and Y. Zhang, "AI fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias," *CoRR*, vol. abs/1810.01943, 2018. [Online]. Available: http://arxiv.org/abs/1810.01943

[36] M. Broussard, *Artificial unintelligence: How computers misunderstand the world*. mit Press, 2018.

[37] C. O'neil, *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown, 2016.

[38] K. Peng, A. Mathur, and A. Narayanan, "The harms that arise and persist after release: An analysis of three dataset life cycles through a thousand papers," in *Responsible Computer Vision Workshop of CVPR*, 2021.