A Framework for Understanding Unintended Consequences of Machine Learning

Harini Suresh
MIT
hsuresh@mit.edu

John V. Guttag
MIT
guttag@mit.edu

Abstract

As machine learning increasingly affects people and society, it is important that we strive for a comprehensive and unified understanding of how and why unwanted consequences arise. For instance, downstream harms to particular groups are often blamed on "biased data," but this concept encompass too many issues to be useful in developing solutions. In this paper, we provide a framework that partitions sources of downstream harm in machine learning into five distinct categories spanning the data generation and machine learning pipeline. We describe how these issues arise, how they are relevant to particular applications, and how they motivate different solutions. In doing so, we aim to facilitate the development of solutions that stem from an understanding of application-specific populations and data generation processes, rather than relying on general claims about what may or may not be "fair."

Introduction

Machine learning (ML) is increasingly used to make decisions that affect people's lives. Typically, ML algorithms operate by learning patterns in historical data and generalizing them to unseen data. As a result, problems with the data or development process can lead to different unintended downstream consequences. In recent years, we have seen several such examples, in contexts from predictive policing (Lum and Isaac 2016) to face recognition (Phillips et al. 2011).

Common rhetoric is that various unwanted consequences of ML algorithms arise in some way from "biased data." In this context, the term "bias" refers to an unintended or potentially harmful property of the data. Data, however, is a product of many factors, from the historical context in which it was generated to the particular forms of measurement error it contains. And it is not just the data that causes problems. The ML pipeline involves a series of choices and practices, from evaluation methodology to model definition, that can lead to unwanted effects. As an ML practitioner working on a new application, it is still not straightforward to identify what problems may be present. Even once identified, it is not clear what the appropriate application- and data-specific solution should be, or how this solution generalizes over factors such as time and geography.

Take the following toy scenario: an engineer building a smile-detection system observes that the system has a higher false negative rate for women. Over the next week, she collects many more images of women, so that the proportions of men and women are now equal, and is happy to see the performance on the female subset improve. Meanwhile, her co-worker has a dataset of job candidates and human-assigned ratings, and wants to build an algorithm for predicting the suitability of a candidate. He notices that women are much less likely to be predicted as suitable candidates than men. Inspired by his colleague's success, he collects many more samples of women, but is dismayed to see that his model's behavior does not change. Why did this happen? The *sources* of the disparate performance in either case were different: In the first case, it arose because of a lack of data on women, and introducing more data solved the issue. In the second case, the use of a proxy label (human assessment of quality) versus the true label (actual qualification) allowed the model to discriminate by gender, and collecting more labelled data from the same distribution did not help.

The contribution of this paper is a new framework and language for partitioning *sources* of downstream harm into five distinct categories. In doing so, we:

1. Provide a consolidated and comprehensive terminology for effectively understanding and connecting work in ML fairness, unpacking broad and/or overloaded terms (e.g. "training data bias").

2. Illustrate that particular solutions make implicit but important assumptions about the data and domain that should be made explicit. We envision future papers being able to state the problem(s) they address in clear, shared terminology, making their framing and assumptions immediately understandable.

3. Facilitate solutions that stem from an understanding of the data generation and analysis processes of a particular application, as opposed to solutions that stem from global assumptions about what it means to be fair.

We note that the categories we define are not mutually exclusive; in fact any one application could suffer from any combination of them. Identifying and characterizing each one as distinct, however, makes them less confusing and easier to tackle.

(a) Data Generation

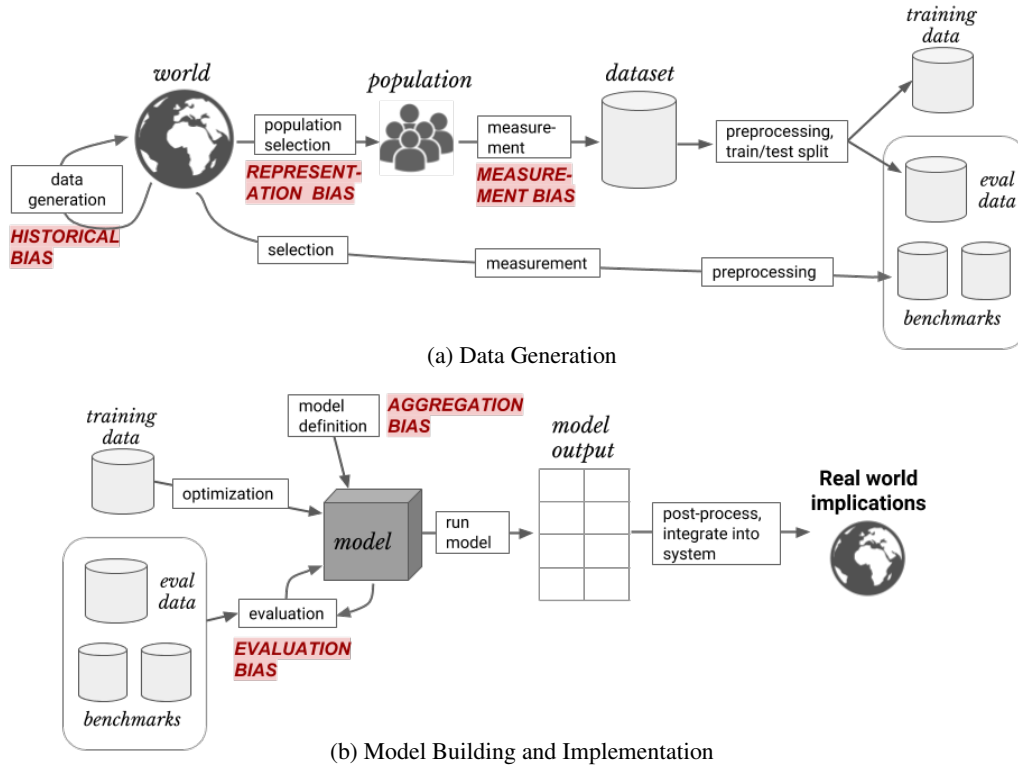

(b) Model Building and Implementation

Figure 1: (a) The data generation process begins with data collection from the world. This process involves both sampling from a population and identifying which features and labels to use. This dataset is split into training and evaluation sets, which are used to develop and evaluate a particular model. Data is also collected (perhaps by a different process) into benchmark datasets. (b) Benchmark data is used to evaluate, compare, and motivate the development of better models. A final model then generates its output, which has some real world manifestation. This process is naturally cyclic, and decisions influenced by models affect the world that exists the next time data is collected or decisions are applied. In red, we indicate where in this pipeline different sources of downstream harm can arise.

A Broad View

In this work, we identify issues that commonly arise in ML applications that lead to some unwanted or societally unfavorable outcome (discussed in more detail below in *Societal Harms*). We argue that analyzing consequences of a particular algorithm should begin with a thorough understanding of the data generation and ML pipeline that led to its output.

The sources of harm we consider arise at different points in such a pipeline (see Figure 1). **Historical bias** is a normative concern with the world as it is; it is a fundamental, structural issue with the first step of the data generation process and can exist even given perfect sampling and feature selection. **Representation bias** arises when defining and sampling from a population, and **measurement bias** arises when subsequently choosing and measuring the particular features of interest. **Evaluation bias** occurs during model iteration and evaluation. **Aggregation bias** arises when flawed assumptions about the population affect model definition. Each of these is described in detail later in the paper.

Knowledge of an application can and should inform the identification of bias sources. Issues that arise in image recognition, for example, are often related to *selection* or *evaluation* bias. Images themselves tend to be objective, but they may not equally represent the entire space of im-

ages that we care about. In data that is affected by human decision-makers, we often see human decisions used as proxies, introducing *measurement* bias. For example, "arrested" is used as a proxy for "crime," or "pain medication prescribed by doctor" is used as a proxy for "patient's pain." Identifying *aggregation* bias usually requires some understanding of meaningful groups and reason to think they are distributed differently. Medical applications, for example, often risk aggregation bias because patients with similar underlying conditions present and progress in different ways. Recognizing *historical bias* requires a retrospective understanding of the application and data generation process over time.

Background

Societal Harms

Barocas et al. (2017) describe a useful framework for thinking about how the negative consequences of automated systems actually manifest, splitting them into *allocative* and *representational* harms. Allocative harms occur when opportunities or resources are withheld from certain groups, while representational harms occur when a system diminishes a particular identity. Consider a search engine that disproportionately displays ads about criminal records when

African American names are searched (Sweeney 2013). If this then leads to racial discrimination against loan applicants, that would be an allocative harm. Even if it does not, however, the perpetuation of racial stereotypes is still a representational harm.

Fairness Definitions and Mitigations

Formalizing Fairness   Many works have gone on to formalize mathematical notions used to measure some of these harms. Typically, these are framed around problems of classification and decision-making, and so address allocative harms. Broadly, they describe some criterion that should be met in order for the algorithm to be considered "fair." They fall into several categories, many of which presuppose a "sensitive attribute" on which examples can be split into groups of interest:

- **Group-Independent Predictions** (Zemel et al. 2013; Feldman et al. 2015; Corbett-Davies et al. 2018) require that the decisions that are made are independent (or conditionally independent) of group membership. For example, the *demographic parity criterion* requires that predictions are uncorrelated with the sensitive attribute.

- **Equal Metrics Across Groups** (Chouldechova 2017; Corbett-Davies et al. 2017) require equal prediction metrics of some sort (this could be accuracy, true positive rates, false positive rates, and so on) across groups. For example, the *equality of opportunity* criterion requires equal true positive rates across groups (Hardt et al. 2016).

- **Individual Fairness** (Johndrow and Lum 2017; Yona and Rothblum 2018) requires that individuals who are similar with respect to the prediction task are treated similarly. The implicit assumption is that there exists an ideal feature space in which to compute similarity, that is reflected or recoverable in the available data. For example, *fairness through awareness* tries to identify a task-specific similarity metric in which individuals who are close according to this metric are also close in outcome space (Dwork et al. 2012).

- **Causal Fairness** (Kilbertus et al. 2017; Nabi and Shpitser 2018) definitions place some requirement on the causal graph that generated the data and outcome. For example, *counterfactual fairness* requires that there is not a causal pathway from a sensitive attribute to the outcome decision (Kusner et al. 2017).

Interested readers are referred to Narayanan (2018) or Verma and Rubin (2018) for a more detailed discussion of different fairness definitions.

Fair Algorithms   Techniques to design "fair" algorithms typically identify a fairness notion of interest and modify the modeling pipeline to satisfy it. Methodologically, they fall broadly into:

- *Data-based methods*, such as data reweighting, e.g., up-weighting examples that align with a particular fairness objective (Hajian and Domingo-Ferrer 2013; Kamiran, Žliobaitė, and Calders 2013; Kamiran and Calders 2012; Calders, Kamiran, and Pechenizkiy 2009).

- *Model-based methods*, such as the addition of regularization terms or constraints that enforce a particular fairness objective during optimization, or representation learning to "mask" out a sensitive variable (Zafar et al. 2017; Zemel et al. 2013; Luong, Ruggieri, and Turini 2011; Berk et al. 2017).

- *Post-hoc methods*, such as identifying different decision thresholds for different groups based on a predicted score, e.g., in order to equalize false positive rates (Hardt et al. 2016; Corbett-Davies and Goel 2018).

For more analysis, Friedler et al. (2019) compare several mitigation techniques on a number of benchmark datasets.

Most existing fairness definitions and algorithms make the implicit assumption that an underlying procedural or statistical notion of fairness can be mathematically defined and operationalized to create a fair system. This assumption often does not address domain-specific societal and historical contexts (Green and Hu 2018). As a result, because applications are so different, these methods have limited scope. For instance, while ensuring group-independent predictions might make sense in hiring (when it is illegal to factor gender or ethnicity into decisions), it would not be appropriate in a medical application where gender and race can play an important role in understanding a patient's symptoms. The framework we provide is inherently application-specific because it stems from identifying the sources of harm throughout the full data generation and ML pipeline.

Related Work

Several works have presented overviews of issues that arise in real datasets. Some of these have been through the lens of a particular domain: Barocas and Selbst (2016) examine various problems that arise in a data-mining pipeline through the lens of American anti-discrimination law, and Danks and London (2017) present a taxonomy of "algorithmic bias," focusing on its manifestation in autonomous systems. This work was extended by Silva and Kenney (2018) to address racial bias in decision-making more generally, though categories in their taxonomy such as "training data bias" encompass many distinct subproblems. Others have focused on a particular subset of the ML pipeline: Calders and Žliobaitė (2013) go over assumptions that are commonly made in the modeling process, and problems that arise because of incorrect labels, skewed sampling procedures, or incomplete data. Friedler, Scheidegger, and Venkatasubramanian (2016) also discuss assumptions about populations that underlie solutions, focusing on how the measured data relates to the desired data.

Our work provides a more comprehensive and unified view on how these issues arise. We aim to provide a shared framework that can encompass many applications while facilitating task-specific social analysis and targeted mitigation.

Five Sources of Bias in ML
Historical Bias
Historical bias arises even if the data is perfectly measured and sampled, if the world *as it is* leads a model to produce

outcomes that are not wanted. For instance, even if we had access to the perfectly-measured feature "crime" in the previous example, it might still reflect historical factors that have led to more crime in poorer neighborhoods. Such a system, even if it reflects the world accurately, can still inflict harm on a population. Considerations of historical bias tend to involve evaluating the representational harm (such as reinforcing a stereotype) to a particular identity group.

Example: image search In 2018, 5% of Fortune 500 CEOs were women (Zarya 2018). Should image search results for "CEO" reflect that number? Ultimately, a variety of stakeholders, including affected members of society, should evaluate the particular harms that this result could cause and make a judgment. This decision may be at odds with the available data even if that data is a perfect reflection of the world. Indeed, Google has recently changed their Image Search results for "CEO" to display a higher proportion of women.

Representation Bias

Representation bias occurs when the certain parts of the input space are underrepresented. A supervised machine learning algorithm aims to learn a function f that minimizes $E_{(x,y)\sim\hat{P}}[J(f(x), y)]$, where $\hat{P}$ is a probability distribution over the input space and J is a loss function, e.g., 0-1 loss for binary classification. If $D(\hat{P}\|P) \geq 0$ where D is a measure of divergence and P is the true data distribution (i.e., $\hat{P}$ is not equal to the true data distribution), this is *selection bias* in the traditional statistical sense.

Representation bias arises if $\hat{P}$ is a distribution that samples too few examples from a particular part of the input space. When we lack data about some part of the input space, the learned mapping f will be more uncertain for new x, y pairs in that area. It is worth noting that even if $D(\hat{P}\|P) = 0$ (i.e., no *selection bias*), representation bias can still occur: if some group is a minority that only makes up 5% of the true distribution, then even sampling from the true data distribution will likely lead to a significantly less robust model for this group.

Representation bias can arise for several reasons, including:

1. **The sampling methods only reach a portion of the population.** For example, datasets collected through smartphone apps can under-represent lower-income or older groups, who are less likely to own smartphones. Similarly, medical data for a particular condition may only be available for the population of patients who were considered serious enough to bring in for further screening.

2. **The population of interest has changed or is distinct from the population used during model training.** Data that is representative of Boston, for example, may not be representative if used to analyze the population of Indianapolis. Similarly, data representative of Boston 30 years ago will likely not reflect today's population.

Example: geographic diversity in image datasets ImageNet is a widely-used image dataset consisting of 1.2 million labeled images (Deng et al. 2009). Approximately 45% of the images in ImageNet were taken in the United States, and the majority of the remaining images are from North America or Western Europe. 1% and 2.1% of the images come from China or India, respectively. Shankar et al. (2017) show that the performance of a classifier trained on ImageNet is significantly worse for several categories (such as "bridegroom") on images that are crowdsourced from under-represented countries such as Pakistan or India versus images from North America and Western Europe.

Measurement Bias

Available, measured data are often *proxies* for some ideal features and labels. For example, arrest rates are often used as a proxy for crime rates. If the measurement process just adds random noise, the model parameters will converge to the those we would expect with the correctly measured features (given enough data). On the other hand, measurement bias often arises because proxies are generated differently across groups (also known as *differential measurement error* (VanderWeele and Hern'an 2012)).

Measurement bias can arise in several ways:

1. **The granularity of data varies across groups.** For example, if a group of factory workers is more stringently or frequently monitored, more errors will be observed J in that group. This can also lead to a feedback loop wherein the group is subject to further monitoring because of the apparent higher rate of mistakes (Barocas and Selbst 2016; Ensign et al. 2017).

2. **The quality of data varies across groups.** Structural discrimination can lead to systematically higher error rates in a certain group. For example, women are more likely to be misdiagnosed or not diagnosed for conditions where self-reported pain is a symptom (in this case "*diagnosed* with condition X" is a biased proxy for "has condition X").

3. **The defined classification task is an oversimplification.** In order to build a supervised ML model, some label to predict must be chosen. Reducing a decision to a single attribute can create a biased proxy label because it only captures a particular aspect of what we really want to measure. Consider the prediction problem of deciding whether a student will be successful (e.g., in a college admissions context). Fully capturing the outcome of 'successful student' in terms of a single measurable attribute is impossible because of its complexity. In cases such as these, algorithm designers resort to some available label such as 'GPA' (Kleinberg et al. 2018), which ignores different indicators of success exhibited by parts of the population.

Example: predictive policing As mentioned previously, in predictive policing applications, the proxy variable "arrest" is often used to measure "crime" or some underlying notion of "riskiness." Because minority communities are often more highly policed and have higher arrest rates, there is a different mapping from crime to arrest for people from these communities. Prior arrests and friend/family arrests were two of many differentially mismeasured proxy variables used in the recidivism risk prediction tool COMPAS

(Angwin et al. 2016), and was a factor that eventually led to higher false positive rates for black versus white defendants. It's worth noting that even such an evaluation is complicated by the proxy label "rearrest" used to measure "recidivism" (Dressel and Farid 2018).

Aggregation Bias

Aggregation bias arises when a one-size-fit-all model is used for groups with different conditional distributions, $p(Y|X)$. Underlying aggregation bias is an assumption that the mapping from inputs to labels is consistent across groups. In reality, this is often not the case. Group membership can be indicative of different backgrounds, cultures or norms, and a given variable can mean something quite different for a person in a different group.

Aggregation bias can lead to a model that is not optimal for any group, or a model that is fit to the dominant population (if combined with representation bias). If there is a non-linear relationship between group membership and outcome, for example, any single linear classifier will have to sacrifice performance on one or both groups. In some cases, incorporating information about group differences into the design of a model can lead to simpler learned functions that improve performance across groups (Dwork et al. 2017; Suresh, Gong, and Guttag 2018).

Example: clinical-aid tools Diabetes patients have known differences in associated complications across ethnicities (Spanakis and Golden 2013). Studies have also suggested that HbA1c levels (widely used to diagnose and monitor diabetes) differ in complex ways across ethnicities and genders (Herman and Cohen 2012). Because these factors have different meanings and importances within different subpopulations, a single model is unlikely to be best-suited for any group in the population even if they are equally represented in the training data.

Evaluation Bias

Evaluation bias occurs when the evaluation and/or benchmark data for an algorithm doesn't represent the target population. A model is optimized on its training data, but its quality is often measured on benchmarks (e.g., UCI datasets (Huang et al. 2007), Faces in the Wild (Dheeru and Karra Taniskidou 2017), ImageNet (Deng et al. 2009)), so a misrepresentative benchmark encourages the development of models that only perform well on a subset of the population.

Evaluation bias ultimately arises because of a need to objectively compare models against each other. Applying different models to some set of external datasets attempts to serve this purpose, but is often extended to make general statements about how good a model is. Such generalizations are often not statistically valid (Salzberg 1997), and can lead to overfitting to a particular benchmark or set of benchmarks. This is especially problematic if the benchmark is not representative. This process is self-fulfilling, as Hand (2006) points out: "Indeed, the more successful the collection is in the sense that more and more people use it for comparative

assessments, the more serious this problem [overfitting to particular benchmarks] will become."

Evaluation bias can be exacerbated by the particular metrics that are used to report performance (both on a model's own test data and on external benchmarks). For example, aggregate measures can hide subgroup underperformance (Suresh, Gong, and Guttag 2018), but such metrics are often used because a single measure makes it straightforward and quick to compare models and make a judgment on which one is "better." Just looking at a single type of metric (e.g., accuracy) can also hide disparities in other types of errors (e.g., false positive rate).

Example: underperformance of commercial facial recognition algorithms Buolamwini and Gebru (2018) point out the drastically worse performance of commercially-used facial analysis algorithms (performing tasks such as gender- or smiling- detection) on dark-skinned females. Looking at some common facial analysis benchmark datasets, it becomes apparent why such algorithms were considered appropriate for use – just 7.4% and 4.4% of the images in benchmark datasets such as Adience and IJB-A are of dark-skinned female faces. Algorithms that underperform on this slice of the population therefore suffer quite little in their evaluation performance on these benchmarks. The algorithms' underperformance was likely due to representation bias in the training data, but the benchmarks failed to discover and penalize this. Since this study, other algorithms have been benchmarked on more balanced face datasets, changing the development process to encourage models that perform well across groups (Ryu, Adam, and Mitchell 2018).
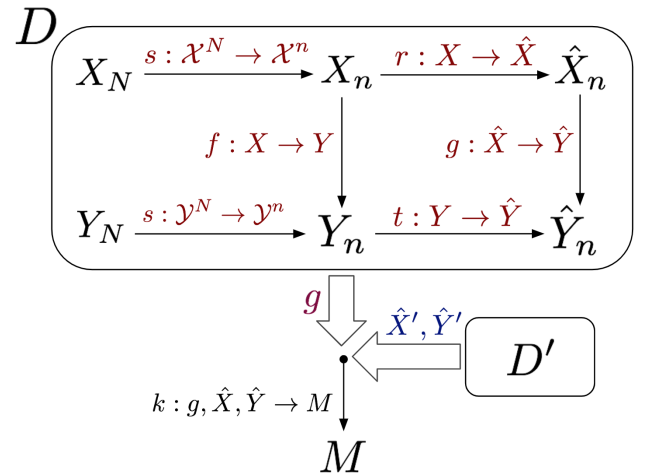
Formalizations and Mitigations

Figure 2: A data generation and ML pipeline viewed as a series of mapping functions. The upper part of the diagram deals with data collection and model building, while the bottom half describes the evaluation process. See the text for a detailed walk-through.

The implications of identifying a particular type of bias are highlighted by abstracting the ML pipeline to a series of

data transformations. Consider the data transformations for a particular dataset (Figure 2). Let X and Y be the ideal, underlying features and labels we wish to capture. The subscript indicates the size of the populations, so X_N indicates the entire population and X_n indicates the smaller population that is actually used, where $s : \mathcal{X}^N \to \mathcal{X}^n$ is the sampling function. $\hat{X}$ and $\hat{Y}$ are the measured and available features and labels that are chosen to build a model, where r and t are the projections from $X \to \hat{X}$ and $Y \to \hat{Y}$ that result in this data. The function $f : X \to Y$ is what we want to learn, but $g : \hat{X} \to \hat{Y}$ is the actual function that is learned. This data transformation sequence can be abstracted into a general process D. Then, k computes some measurement of success M for g on data $\hat{X}', \hat{Y}'$ that can come from the same pipeline or a different one (D' in Figure 2).

Measurement and historical biases are issues with how features and labels are projected, i.e., how r and t are instantiated. Therefore, we can see that solutions that try to adjust s (e.g., collecting more data that then undergoes the same transformation to $\hat{X}$) will likely be ineffective. Representation bias stems from a problem with s, the sampling function; methods that then adjust r or t (e.g., choosing different features or labels) or g (e.g., changing the objective function) may be misguided. Importantly, solutions that *do* address representation bias by adjusting s implicitly assume that r and t are acceptable (e.g., they are something close to the identity function) and therefore, improving s will mitigate the issue. Benchmark and aggregation bias are discussed in more detail in the case studies below.

Case Study 1: Mitigating Aggregation Bias

Aggregation bias is a limitation on the learned function g (e.g., a linear parameterization) that stems from an assumption about the homogeneity of $p(\hat{Y}|\hat{X})$. Ultimately, this results in a g that is disproportionately worse for some group(s). Addressing limitations of g can be achieved by either 1. adjusting g to better model the data complexities, or 2. transforming the data such that g is now better suited to it.

Methods that adjust g include **coupled learning methods**, such as multitask learning, that parameterize different groups differently in the model definition and facilitate learning multiple simpler functions that take into account group differences (Dwork et al. 2017; Suresh, Gong, and Guttag 2018; Oneto et al. 2018).

To transform the data, we need to change r or t. **Fair representation learning** involves projecting data into a space (i.e., coming up with a new mapping $r : X \to \hat{X}$) where examples that are similar with respect to the prediction task are close to each other in feature space (i.e., projecting into a space where $p(\hat{Y}|\hat{X})$ is the same across groups), and then learning g (Zemel et al. 2013; Louizos et al. 2015). This space aims to capture some true underlying features that may manifest differently across groups. Note that solutions such as *anti-classification* (Corbett-Davies et al. 2018) or *fairness through unawareness* (Gajane and Pechenizkiy 2017) that make predictions independently of group membership do not address aggregation bias.

Case Study 2: Mitigating Evaluation Bias

Evaluation bias is an issue with M, a measurement of the quality of the learned function, g. If we trace the inputs to M, we can see that addressing M would involve 1) redefining k (i.e., the function that computes evaluation metrics) and/or 2) adjusting $\hat{X}$, $\hat{Y}$ (data and labels from a separate benchmark dataset).

Improving k involves making it more *comprehensive* and *granular*. The *granularity* of k can be improved with **subgroup evaluation** that compares per-group metrics as well as aggregate measures that weight groups equally (Buolamwini and Gebru 2018). Deciding what groups to use is often application-specific and requires intersectional analysis and privacy consideration; see Mitchell et al. (2018) for a more in-depth discussion. **Multiple metrics and confidence intervals** improve the *comprehensiveness* of the evaluation. Choosing the metrics of interest should involve domain specialists and affected populations that understand the usage and consequences of the model. In a predictive policing application, for example, law enforcement may prioritize a low false negative rate (not missing any high-risk people) while affected communities may value a low false positive rate (not being mistakenly classified as high-risk). Section 4.4 of Mitchell et al. (2018) further discusses different metrics.

Issues with $\hat{X}'$ and $\hat{Y}'$ stem from an unrepresentative sampling function s'. Improving s' may involve **targeted data augmentation** (e.g., SMOTE) to populate parts of the data distribution that are underrepresented (Iosifidis and Ntoutsi 2018; Chawla et al. 2002).

Conclusions

We provide a framework for understanding "bias" in ML at a level of abstraction that we hope will facilitate productive communication and development of solutions. Terms such as "training data bias" are too broad to be useful, and context-specific fixes don't have the shared terminology to generalize and communicate the problem to a wider audience. We envision future work being able to state upfront which particular type of bias they are addressing, making it immediately clear what problem they are trying to solve and making assumptions explicit rather than implicit.

By framing sources of downstream harm through the data generation process, we encourage application-appropriate solutions rather than relying on broad notions of what is fair. Fairness is not one-size-fits-all; knowledge of an application and engagement with its stakeholders should inform the identification of these sources.

Finally, we illustrate that there are important choices being made throughout the larger data generation and ML pipeline that extend far beyond model building. In practice, ML is an iterative process with a long and complicated feedback loop. We highlight problems that manifest through this loop, from historical context to the process of evaluating models and benchmarking them against each other.

References

[Angwin et al. 2016] Angwin, J.; Larson, J.; Mattu, S.; and Kirchner, L. 2016. Machine bias. *ProPublica, May* 23.

[Barocas and Selbst 2016] Barocas, S., and Selbst, A. D. 2016. Big data's disparate impact. *Cal. L. Rev.* 104:671.

[Barocas et al. 2017] Barocas, S.; Crawford, K.; Shapiro, A.; and Wallach, H. 2017. The problem with bias: from allocative to representational harms in machine learning. *Special Interest Group for Computing, Information and Society (SIGCIS).*

[Berk et al. 2017] Berk, R.; Heidari, H.; Jabbari, S.; Joseph, M.; Kearns, M.; Morgenstern, J.; Neel, S.; and Roth, A. 2017. A convex framework for fair regression. *arXiv preprint arXiv:1706.02409.*

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

[Calders and Žliobaitė 2013] Calders, T., and Žliobaitė, I. 2013. Why unbiased computational processes can lead to discriminative decision procedures. In *Discrimination and privacy in the information society*. Springer. 43–57.

[Calders, Kamiran, and Pechenizkiy 2009] Calders, T.; Kamiran, F.; and Pechenizkiy, M. 2009. Building classifiers with independency constraints. In *Data mining workshops, 2009. ICDMW'09. IEEE international conference on*, 13–18. IEEE.

[Chawla et al. 2002] Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; and Kegelmeyer, W. P. 2002. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 16:321–357.

[Chouldechova 2017] Chouldechova, A. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* 5(2):153–163.

[Corbett-Davies and Goel 2018] Corbett-Davies, S., and Goel, S. 2018. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023.*

[Corbett-Davies et al. 2017] Corbett-Davies, S.; Pierson, E.; Feller, A.; Goel, S.; and Huq, A. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 797–806. ACM.

[Corbett-Davies et al. 2018] Corbett-Davies, S.; Goel, S.; Morgenstern, J.; and Cummings, R. 2018. Defining and designing fair algorithms. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, 705–705. ACM.

[Danks and London 2017] Danks, D., and London, A. J. 2017. Algorithmic bias in autonomous systems. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, 4691–4697.

[Deng et al. 2009] Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*.

[Dheeru and Karra Taniskidou 2017] Dheeru, D., and Karra Taniskidou, E. 2017. UCI machine learning repository.

[Dressel and Farid 2018] Dressel, J., and Farid, H. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science advances* 4(1):eaao5580.

[Dwork et al. 2012] Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, 214–226. ACM.

[Dwork et al. 2017] Dwork, C.; Immorlica, N.; Kalai, A. T.; and Leiserson, M. 2017. Decoupled classifiers for fair and efficient machine learning. *arXiv preprint arXiv:1707.06613.*

[Ensign et al. 2017] Ensign, D.; Friedler, S. A.; Neville, S.; Scheidegger, C. E.; and Venkatasubramanian, S. 2017. Runaway feedback loops in predictive policing. *CoRR* abs/1706.09847.

[Feldman et al. 2015] Feldman, M.; Friedler, S. A.; Moeller, J.; Scheidegger, C.; and Venkatasubramanian, S. 2015. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 259–268. ACM.

[Friedler et al. 2019] Friedler; Scheidegger; Venkatasubramanian; Choudhary; Hamilton; and Roth. 2019. A comparative study of fairness-enhancing interventions in machine learning. In *ACM Conference on Fairness, Accountability and Transparency (FAT*)*. ACM.

[Friedler, Scheidegger, and Venkatasubramanian 2016] Friedler, S. A.; Scheidegger, C.; and Venkatasubramanian, S. 2016. On the (im) possibility of fairness. *arXiv preprint arXiv:1609.07236.*

[Gajane and Pechenizkiy 2017] Gajane, P., and Pechenizkiy, M. 2017. On formalizing fairness in prediction with machine learning. *arXiv preprint arXiv:1710.03184.*

[Green and Hu 2018] Green, B., and Hu, L. 2018. The myth in the methodology: towards a recontextualization of fairness in machine learning. ICML.

[Hajian and Domingo-Ferrer 2013] Hajian, S., and Domingo-Ferrer, J. 2013. A methodology for direct and indirect discrimination prevention in data mining. *IEEE transactions on knowledge and data engineering* 25(7):1445–1459.

[Hand 2006] Hand, D. J. 2006. Classifier technology and the illusion of progress. *Statistical science* 1–14.

[Hardt et al. 2016] Hardt, M.; Price, E.; Srebro, N.; et al. 2016. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, 3315–3323.

[Herman and Cohen 2012] Herman, W. H., and Cohen, R. M. 2012. Racial and ethnic differences in the relationship between hba1c and blood glucose: implications for the diagnosis of diabetes. *The Journal of Clinical Endocrinology & Metabolism* 97(4):1067–1072.

[Huang et al. 2007] Huang, G. B.; Ramesh, M.; Berg, T.; and Learned-Miller, E. 2007. Labeled faces in the wild: A

database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst.

[Iosifidis and Ntoutsi 2018] Iosifidis, V., and Ntoutsi, E. 2018. Dealing with bias via data augmentation in supervised learning scenarios. 24–29.

[Johndrow and Lum 2017] Johndrow, J. E., and Lum, K. 2017. An algorithm for removing sensitive information: application to race-independent recidivism prediction. *arXiv preprint arXiv:1703.04957*.

[Kamiran and Calders 2012] Kamiran, F., and Calders, T. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems* 33(1):1–33.

[Kamiran, Žliobaitė, and Calders 2013] Kamiran, F.; Žliobaitė, I.; and Calders, T. 2013. Quantifying explainable discrimination and removing illegal discrimination in automated decision making. *Knowledge and information systems* 35(3):613–644.

[Kilbertus et al. 2017] Kilbertus, N.; Carulla, M. R.; Parascandolo, G.; Hardt, M.; Janzing, D.; and Schölkopf, B. 2017. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems*, 656–666.

[Kleinberg et al. 2018] Kleinberg, J.; Ludwig, J.; Mullainathan, S.; and Rambachan, A. 2018. Algorithmic fairness. In *AEA Papers and Proceedings*, volume 108, 22–27.

[Kusner et al. 2017] Kusner, M. J.; Loftus, J.; Russell, C.; and Silva, R. 2017. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, 4066–4076.

[Louizos et al. 2015] Louizos, C.; Swersky, K.; Li, Y.; Welling, M.; and Zemel, R. 2015. The variational fair autoencoder. *arXiv preprint arXiv:1511.00830*.

[Lum and Isaac 2016] Lum, K., and Isaac, W. 2016. To predict and serve? *Significance* 13(5):14–19.

[Luong, Ruggieri, and Turini 2011] Luong, B. T.; Ruggieri, S.; and Turini, F. 2011. k-nn as an implementation of situation testing for discrimination discovery and prevention. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 502–510. ACM.

[Mitchell et al. 2018] Mitchell, M.; Wu, S.; Zaldivar, A.; Barnes, P.; Vasserman, L.; Hutchinson, B.; Spitzer, E.; Raji, I. D.; and Gebru, T. 2018. Model cards for model reporting. *arXiv preprint arXiv:1810.03993*.

[Nabi and Shpitser 2018] Nabi, R., and Shpitser, I. 2018. Fair inference on outcomes. In *Proceedings of the... AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence*, volume 2018, 1931. NIH Public Access.

[Narayanan 2018] Narayanan, A. 2018. Fat* tutorial: 21 fairness definitions and their politics. *New York, NY, USA*.

[Oneto et al. 2018] Oneto, L.; Donini, M.; Elders, A.; and Pontil, M. 2018. Taking advantage of multitask learning for fair classification. *arXiv preprint arXiv:1810.08683*.

[Phillips et al. 2011] Phillips, P. J.; Jiang, F.; Narvekar, A.; Ayyad, J.; and O'Toole, A. J. 2011. An other-race effect for face recognition algorithms. *ACM Transactions on Applied Perception (TAP)* 8(2):14.

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

[Salzberg 1997] Salzberg, S. L. 1997. On comparing classifiers: Pitfalls to avoid and a recommended approach. *Data mining and knowledge discovery* 1(3):317–328.

[Shankar et al. 2017] Shankar, S.; Halpern, Y.; Breck, E.; Atwood, J.; Wilson, J.; and Sculley, D. 2017. No classification without representation: Assessing geodiversity issues in open data sets for the developing world. *arXiv preprint arXiv:1711.08536*.

[Silva and Kenney 2018] Silva, S., and Kenney, M. 2018. Algorithms, platforms, and ethnic bias: An integrative essay.

[Spanakis and Golden 2013] Spanakis, E. K., and Golden, S. H. 2013. Race/ethnic difference in diabetes and diabetic complications. *Current diabetes reports* 13(6):814–823.

[Suresh, Gong, and Guttag 2018] Suresh, H.; Gong, J. J.; and Guttag, J. V. 2018. Learning tasks for multitask learning: Heterogenous patient populations in the icu. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM.

[Sweeney 2013] Sweeney, L. 2013. Discrimination in online ad delivery. *Queue* 11(3):10.

[VanderWeele and Hern'an 2012] VanderWeele, T. J., and Hern'an, M. A. 2012. Results on differential and dependent measurement error of the exposure and the outcome using signed directed acyclic graphs. *American journal of epidemiology* 175(12):1303–1310.

[Verma and Rubin 2018] Verma, S., and Rubin, J. 2018. Fairness definitions explained.

[Yona and Rothblum 2018] Yona, G., and Rothblum, G. 2018. Probably approximately metric-fair learning. In *International Conference on Machine Learning*, 5666–5674.

[Zafar et al. 2017] Zafar, M. B.; Valera, I.; Gomez Rodriguez, M.; and Gummadi, K. P. 2017. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, 1171–1180. International World Wide Web Conferences Steering Committee.

[Zarya 2018] Zarya, V. 2018. The share of female ceos in the fortune 500 dropped by 25% in 2018. *Fortune*.

[Zemel et al. 2013] Zemel, R.; Wu, Y.; Swersky, K.; Pitassi, T.; and Dwork, C. 2013. Learning fair representations. In *International Conference on Machine Learning*, 325–333.