MIT Case Studies in Social and Ethical Responsibilities of Computing • Summer 2021

Understanding Potential Sources of Harm throughout the Machine Learning Life Cycle

Harini Suresh[1], John Guttag[1]

[1]Computer Science and Artificial Intelligence Laboratory, MIT

Published on: Aug 10, 2021

DOI: 10.21428/2c646de5.c16a07bb

ABSTRACT

As machine learning (ML) increasingly affects people and society, awareness of its potential unwanted consequences has also grown. To anticipate, prevent, and mitigate undesirable downstream consequences, it is critical that we understand when and how harm might be introduced throughout the ML life cycle. In this case study, we provide a framework that identifies seven distinct potential sources of downstream harm in machine learning, spanning data collection, development, and deployment. We describe how these issues arise, how they are relevant to particular applications, and how they motivate different mitigations.

Keywords: fairness in machine learning; societal implications of machine learning; algorithmic bias; AI ethics

Harini Suresh
Department of Electrical Engineering and Computer Science, and Computer Science and Artificial Intelligence Laboratory, MIT

John Guttag
Department of Electrical Engineering and Computer Science, and Computer Science and Artificial Intelligence Laboratory, MIT

Learning Objectives

- Learners will understand the various stages of the machine learning (ML) life cycle, and consider the implications of decisions made at each stage.
- Learners will begin to think about data as the product of a complex human-driven process, as opposed to a static, objective artifact.
- Learners will be able to approach related literature or current events with a clearer understanding of why and how machine learning systems can cause harm.

1. Introduction

Machine learning (ML) is increasingly used to make decisions that affect people's lives. Typically, ML algorithms operate by learning models from existing data and generalizing them to unseen data. As a result, problems can arise during the data collection, model development, and deployment processes that can lead to distinct types of harmful downstream consequences. In recent years, we have seen examples in diverse contexts such as facial analysis (e.g., where publicly available algorithms performed significantly worse on dark-skinned women) and pretrial risk assessment of defendants in the criminal justice system (e.g., where a deployed algorithm was more likely to incorrectly predict Black defendants as being high risk).[1]

A common refrain is that undesirable behaviors of ML systems happen when “data is biased.” Indeed, a recent comment by a prominent ML researcher to this end set off a heated debate—not necessarily because the statement “data is biased” is *false*, but because it treats data as a static artifact divorced from the process that produced it.[2] This process is long and complex, grounded in historical context and driven by human choices and norms. Understanding the implications of each stage in the data-generation process can reveal more direct and meaningful ways to prevent or address harmful downstream consequences that can be masked by overly broad terms like “biased data.”

Moreover, it is important to acknowledge that not all problems should be blamed on the data. The ML pipeline involves a series of choices and practices, from model definition to user interfaces used upon deployment. Each stage involves decisions that can lead to undesirable effects. For an ML practitioner working on a new system, it is not straightforward to identify if and how problems might arise. Even once identified, it is not clear what the appropriate application- and data-specific mitigations might be, or how they might generalize over factors such as time and geography.

Consider the following simplified scenario: a medical researcher wants to build a model to help detect whether someone is having a heart attack. She trains the model on medical records from a subset of prior patients at a hospital, along with labels indicating if and when they suffered a heart attack. She observes that the system has a higher false negative rate for women (that is, the model is more likely to miss cases of heart attacks in women), so she hypothesizes that the model was not able to effectively learn the signs of heart attacks in women because of a lack of such examples. She seeks out additional data representing women who experienced heart attacks to augment the data set, retrains the model, and observes that the performance for

female patients improves. Meanwhile, a coworker hiring new lab technicians tries to build an algorithm for predicting the suitability of a candidate from their resume along with human-assigned ratings. He notices that women are much less likely to be predicted as suitable candidates than men. Like his colleague, he tries to collect many more samples of women to add to the data set, but is disappointed to see that the model's behavior does not change. Why did this happen? The *sources* of the disparate performance were different. In the first case, it arose because of a lack of data on women, and introducing more data was helpful. In the second case, using human assessment of quality as a label to estimate true qualification allowed the model to discriminate by gender, and collecting more labeled data from the same distribution did not help.

This case study provides a framework and vocabulary for understanding distinct *sources* of downstream harm from ML systems. We demonstrate how issues can arise at different stages of the ML life cycle, and provide corresponding terminology that avoids overly broad and overloaded terms. Doing so begins to illustrate that existing discourse and literature often makes implicit but important assumptions about the data and domain that should be made explicit. Finally, this more precise framework can begin to facilitate mitigations that stem from an understanding of the data generation and development processes of a particular application, as opposed to solutions that stem from global assumptions about what it means to be fair.

Throughout the case study, we refer to the concept of “harm” or “negative consequences” caused by ML systems. Barocas and colleagues provide a useful framework for thinking about how these consequences actually manifest, splitting them into *allocative* harms (when opportunities or resources are withheld from certain people or groups) and *representational* harms (when certain people or groups are stigmatized or stereotyped).[3] For example, algorithms that determine whether someone is offered a loan or a job risk inflicting allocative harm.[4] This is typically the type of harm that we think and hear about, because it can be measured and is more commonly recognized as harmful. However, even if they do not directly withhold resources or opportunities, systems can still cause representational harm, such as language models that encode and replicate stereotypes.

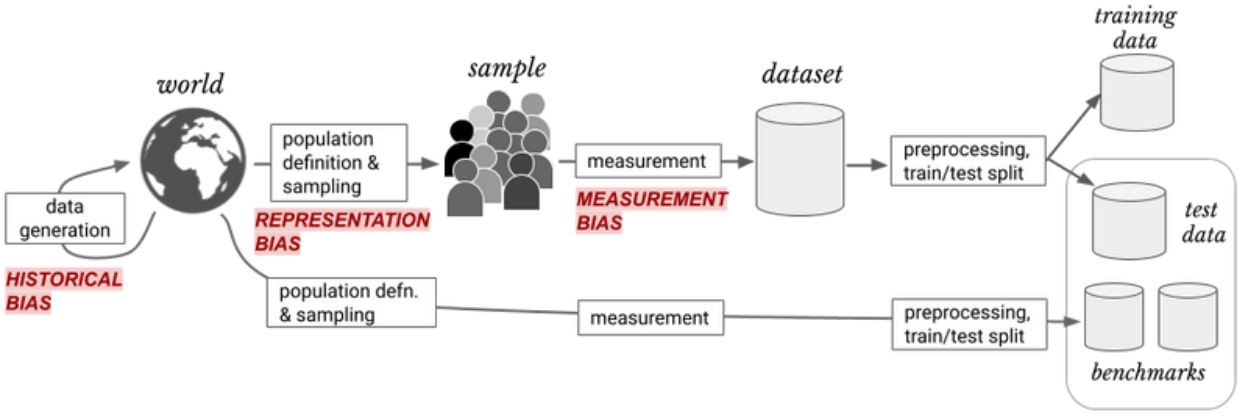
Section 2 follows with a brief overview of the ML pipeline that will be useful background information as we refer to different parts of this process. Section 3 details each source of harm in more depth with examples, and Section 4 is a brief conclusion.

We also include an appendix that provides a more rigorous presentation of our framework for formalizing and mitigating the issues we describe.

2. Machine Learning Overview

Machine learning is a type of statistical inference that learns, from existing data, a function that can be generalized to new, unseen data. We can find ML algorithms all around us: making personalized Netflix or YouTube recommendations, powering Siri's stilted conversation, providing live transcriptions on our video calls, auto-tagging the people in our photos, deciding whether we are offered job interviews, or approving (or not) tests at the doctor's office. In each of these examples, an ML algorithm has found patterns in a (usually massive) data set, and is applying that knowledge to make a prediction about new data points (which might be your photos, medical records, resume, and so on).

In this section, we briefly describe the typical life cycle of an ML system. We will explain each step generally, as well as how it might occur in a running hypothetical example: a machine learning–based loan-approval system. In the running example, we describe each step as it typically happens (not necessarily as it ideally *should*). In the next section, we analyze the implications of each step and problems that may be introduced. Figure 1 depicts these steps.



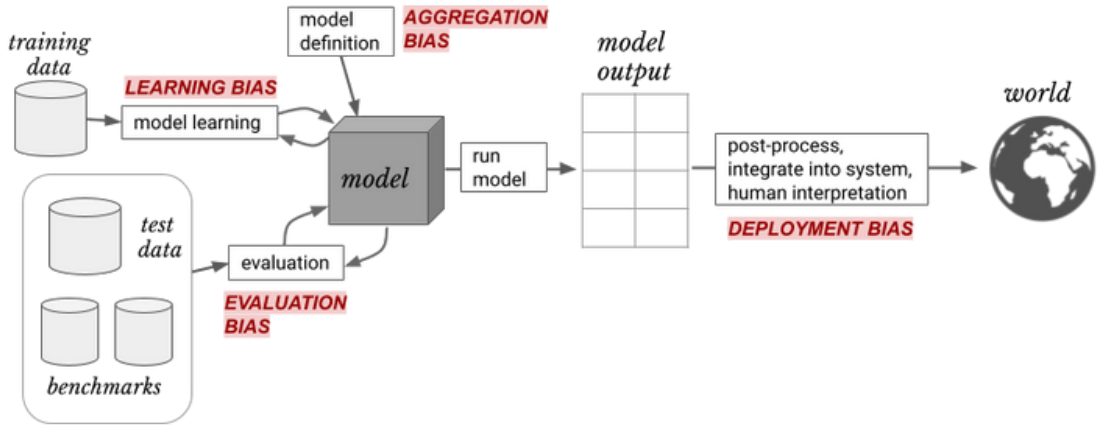


Figure 1. (*Top*) The data generation process begins with data collection. This process involves defining a target population and sampling from it, as well as identifying and measuring features and labels. This data set is split into training and test sets. Data is also collected (perhaps by a different process) into benchmark data sets. (*Bottom*) A model is defined and optimized on the training data. Test and benchmark data are used to evaluate it, and the final model is then integrated into a real-world context. This process is naturally cyclic, and decisions influenced by models affect the state of 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 might arise.

2.1. Data Collection

Before any analysis or learning happens, data must first be collected. Compiling a data set involves identifying a *target population* (of people or things), as well as defining and measuring *features* and *labels* from it. Typically, it is not feasible to include the entire target population, and instead, features and labels are sampled from a subset of it. (We refer to this subset as the *development sample*.) Often, ML practitioners use existing data sets rather than going through the data collection process.

Example. For the loan-approval system, a team in charge of data collection could choose the target population to be people who live in the state in which the system will be used, people who have previously applied for loans, people with credit cards, and so on. The particular sample that ends up in the data set will be a subset of this target population and will depend upon the sampling method (e.g., sourcing information from public records or surveying people). There is also the question of what to actually measure or collect about these people: perhaps things like their debt history, the number of credit cards they have, their income, their occupation, and so on. Some of these things will be chosen to serve as labels: for example, information about whether the person received or paid back a loan in the past.

2.2. Data Preparation

Depending on the data modality and task, different types of preprocessing may be applied to the data set before using it. Data sets are usually split into *training data* used during model development, and *test data* used during model evaluation. Part of the training data may be used as *validation data*.

Example. For the loan-approval system, preprocessing might involve dealing with missing data (e.g., imputing missing credit history values via interpolation), simplifying the feature space (e.g., grouping occupations in broader categories like "physician" rather than encoding detailed specialties), or normalizing continuous measurements (e.g., scaling income so it lies on a 0-to-1 scale). If a resulting data set included 1,000 examples (e.g., data collected from 1,000 people), 600 examples might be used for training, 100 as a validation set during training, and 300 for postdevelopment testing.

2.3. Model Development

Models are then built using the training data (not including the held-out validation data). Typically, models are trained to optimize a specified *objective*, such as minimizing the mean squared error between its predictions and the actual labels. A number of different model types, hyperparameters, and optimization methods may be tested out at this point; usually these different configurations are compared based on their performance on the validation data, and the best one chosen.

Example. The team developing the loan-approval model would first need to instantiate a particular model (e.g., a dense, feed-forward neural network) and define an objective function (e.g., minimizing the cross-entropy loss between the model's predictions and the label defined in the training data). Then, in the optimization process, the model will try to learn a function that goes from the inputs (e.g., income, occupation, etc.) to the output (e.g., whether the person paid back a previous loan). They might also train a number of different models (e.g., with varying architectures or training procedures) and choose the one that performs best on the validation set.

2.4. Model Evaluation

After the final model is chosen, the performance of the model on the test data is reported. The test data is not used before this step, to ensure that the model's performance is a true representation of how it performs on unseen data. Aside from the test data, other available data sets—also called *benchmark data sets*—may be used to demonstrate model robustness or to enable comparison to other existing methods.

The particular *performance metric(s)* used during evaluation are chosen based on the task and data characteristics.

Example. Here, the model developed in the previous step would be evaluated by its performance on the test set. There might be several performance metrics to consider—for example, applicants might be concerned with false negatives (i.e., being denied a loan when they actually are deserving), while lenders might care more about false positives (i.e., recommending loans to people who don't pay them back). In addition, the model might be evaluated on existing data sets used for similar tasks (e.g., the data set from the U.S. Small Business Association described by Min Li and collaborators).[5]

2.5. Model Postprocessing

Once a model has been trained, there are various postprocessing steps that may needed. For example, if the output of a model performing binary classification is a probability, but the desired output to display to users is a categorical answer, there remains a choice of what threshold(s) to use to round the probability to a hard classification.

Example. The resulting model for predicting loan approval likely outputs a continuous score between 0 and 1. The team might choose to transform this score into discrete buckets (e.g., low risk of defaulting, unsure, high risk of defaulting) or a binary recommendation (e.g., should/should not receive a loan).

2.6. Model Deployment

There are many steps that arise in deploying a model to a real-world setting. For example, the model may need to be changed based on requirements for explainability or apparent consistency of results, or there may need to be built-in mechanisms to integrate real-time feedback. Importantly, there is no guarantee that the population a model sees as input after it is deployed (here, we will refer to this as the *use population*) looks the same as the population in the development sample.

Example. In order to deploy the loan-approval system, the team will likely need to develop a user interface that displays the result and the recommended action. They might need to develop different visualizations of the model's reasoning and results for lenders, applicants, regulatory agencies, or other relevant stakeholders. And they may need to incorporate mechanisms for applicants to seek recourse if they believe the model recommendation was inaccurate or discriminatory.

3. Seven Sources of Harm in ML

In this section, we will go into more depth on each potential source of harm. Each subsection will detail where and how in the ML pipeline problems might arise, as well as a characteristic example. These categories are not mutually exclusive; however, identifying and characterizing each one as distinct makes them less confusing and easier to tackle.

3.1. Historical Bias

Historical bias arises even if data is perfectly measured and sampled, if the world *as it is* or *was* leads to a model that produces harmful outcomes. Such a system, even if it reflects the world accurately, can still inflict harm on a population. Considerations of historical bias often involve evaluating the representational harm (such as reinforcing a stereotype) to a particular group.

3.1.1. Example: Word Embeddings

Word embeddings are learned vector representations of words that encode semantic meaning, and are widely used for natural language processing (NLP) applications. Recent research has shown that word embeddings, which are learned from large corpora of text (e.g., Google news, web pages, Wikipedia), reflect human biases. One such study demonstrates that word embeddings reflect real-world biases about women and ethnic minorities, and that an embedding model trained on data from a particular decade reflects the biases of that time.[6] For example, gendered occupation words like “nurse” or “engineer” are highly associated with words that represent women or men, respectively. A range of NLP applications (e.g., chatbots, machine translation, speech recognition) are built using these types of word embeddings, and as a result can encode and reinforce harmful stereotypes.

3.2. Representation Bias

Representation bias occurs when the development sample underrepresents some part of the population, and subsequently fails to generalize well for a subset of the use population. Representation bias can arise in several ways:

1. **When defining the target population, if it does not reflect the use population**. 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 thirty years ago will likely not reflect today’s population.

2. **When defining the target population, if it contains underrepresented groups**. Say the target population for a particular medical data set is defined to be adults aged eighteen to forty. There are minority groups within this population: for example, people who are pregnant may make up only 5 percent of the target population. Even if we sample perfectly, and even if the use population is the same (adults eighteen to forty), the model will likely be less robust for those 5 percent of pregnant people because it has less data to learn from.

3. **When sampling from the target population, if the sampling method is limited or uneven**. For example, the target population for modeling an infectious disease might be all adults, but medical data may be available only for the sample of people who were considered serious enough to bring in for further screening. As a result, the development sample will represent a skewed subset of the target population. In statistics, this is typically referred to as *sampling bias*.

3.2.1. Example: Geographic Diversity in Image Data Sets

ImageNet is a widely used image data set consisting of 1.2 million labeled images.[7] ImageNet is intended to be used widely (i.e., its target population is "all natural images"). However, ImageNet does not evenly sample from this target population; instead, approximately 45 percent 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. Only 1 percent and 2.1 percent of the images come from China and India, respectively. As a result, Shreya Shankar and coauthors have demonstrated that the performance of a classifier trained on ImageNet is significantly worse at classifying images containing certain objects or people (such as "bridegroom") when the images come from undersampled countries such as Pakistan or India.[8]

3.3. Measurement Bias

Measurement bias occurs when choosing, collecting, or computing features and labels to use in a prediction problem. Typically, a feature or label is a *proxy* (a concrete measurement) chosen to approximate some *construct* (an idea or concept) that is not directly encoded or observable. For example, "creditworthiness" is an abstract construct that is often operationalized with a measurable proxy like a credit score. Proxies become problematic when they are poor reflections of the target construct or are generated differently across groups, which can occur when:

1. **The proxy is an oversimplification of a more complex construct**. 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 may resort to a single available label such as "grade-point average" (GPA), which ignores different indicators of success present in different parts of the population.[9]

2. **The method of measurement varies across groups**. For example, consider factory workers at several different locations who are monitored to count the number of errors that occur (i.e., observed number of errors is being used as a proxy for work quality). If one location is monitored much more stringently or frequently, there will be more errors observed for 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.[10]

3. **The accuracy of measurement varies across groups**. For example, in medical applications, "*diagnosed* with condition *X*" is often used as a proxy for "has condition *X*." However, structural discrimination can lead to systematically higher rates of misdiagnosis or underdiagnosis in certain groups.[11] For example, there are both gender and racial disparities in diagnoses for conditions involving pain assessment.[12]

3.3.1. Example: Risk Assessments in the Criminal Justice System

Risk assessments have been deployed at several points within criminal justice settings.[13] For example, risk assessments such as Northpointe's COMPAS predict the likelihood that a defendant will reoffend, and may be used by judges or parole officers to make decisions around pretrial release.[14] The data for models like these often include proxy variables such as "arrest" to measure "crime" or some underlying notion of "riskiness." Because minority communities are more highly policed, this proxy is *differentially mismeasured*—there is a different mapping from crime to arrest for people from these communities. Many of the other features used in COMPAS (e.g., "rearrest" to measure "recidivism") were also differentially measured proxies.[15] The resulting model had a significantly higher false positive rate for Black defendants versus white defendants, that is, the model was more likely to predict that Black defendants were at a high risk of reoffending when they actually were not.

3.4. Aggregation Bias

Aggregation bias arises when a one-size-fits-all model is used for data in which there are underlying groups or types of examples that should be considered differently. Underlying aggregation bias is an assumption that the mapping from inputs to labels is

consistent across subsets of the data. In reality, this is often not the case. A particular data set might represent people or groups with different backgrounds, cultures, or norms, and a given variable can mean something quite different across them. Aggregation bias can lead to a model that is not optimal for any group, or a model that is fit to the dominant population (e.g., if there is also representation bias).

3.4.1. Example: Social Media Analysis

Desmond Patton and colleagues describe analyzing Twitter posts of gang-involved youth in Chicago.[16] By hiring domain experts from the community to interpret and annotate tweets, they were able to identify shortcomings of more general, noncontext-specific NLP tools. For example, certain emojis or hashtags convey particular meanings that a nonspecific model trained on all Twitter data would miss. In other cases, words or phrases that might convey aggression elsewhere are actually lyrics from a local rapper.[17] Ignoring this group-specific context in favor of a single, more general model built for all social media data would likely lead to harmful misclassifications of the tweets from this population.

3.5. Learning Bias

Learning bias arises when modeling choices amplify performance disparities across different examples in the data.[18] For example, an important modeling choice is the objective function that an ML algorithm learns to optimize during training. Typically, these functions encode some measure of accuracy on the task (e.g., cross-entropy loss for classification problems or mean squared error for regression problems). However, issues can arise when prioritizing one objective (e.g., overall accuracy) damages another (e.g., disparate impact).[19] For example, minimizing cross-entropy loss when building a classifier might inadvertently lead to a model with more false positives than might be desirable in many contexts.

3.5.1. Example: Optimizing for Privacy or Compactness

Recent work has explored training models that maintain *differential privacy* (i.e., preventing them from inadvertently revealing excessive identifying information about the training examples during use). However, Eugene Bagdasaryan and colleagues show that differentially private training, while improving privacy, reduces the influence of underrepresented data on the model, and subsequently leads to a model with worse performance on that data (as compared to a model without differentially private training).[20] Similarly, Sara Hooker and collaborators have demonstrated how prioritizing compact models (e.g., with methods such as *pruning*) can amplify

performance disparities on data with underrepresented attributes.[21] This happens because, given limited capacity, the model learns to preserve information about the most frequent features.

3.6. Evaluation Bias

Evaluation bias occurs when the benchmark data used for a particular task does not represent the use population. A model is optimized on its training data, but its quality is often measured on benchmarks (e.g., UCI datasets, Faces in the Wild, ImageNet). This issue operates at a broader scale than other sources of bias: a misrepresentative benchmark encourages the development and deployment of models that perform well only on the subset of the data represented by the benchmark data.

Evaluation bias ultimately arises because of a desire to quantitatively compare models against each other. Applying different models to a set of external data sets 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, and can lead to overfitting to a particular benchmark.[22] This is especially problematic if the benchmark suffers from historical, representation, or measurement bias.

Evaluation bias can also be exacerbated by the choice of metrics that are used to report performance. For example, aggregate measures can hide subgroup underperformance, but these singular measures are often used because they make it more straightforward to compare models and make a judgment about which one is "better." Just looking at one type of metric (e.g., accuracy) can also hide disparities in other types of errors (e.g., false-negative rate).

3.6.1. Example: Commercial Facial Analysis Tools

Joy Buolamwini and Timnit Gebru point out the drastically worse performance of commercial facial analysis algorithms (performing tasks such as gender or smiling detection) on images of dark-skinned women.[23] Images of dark-skinned women comprise only 7.4 percent and 4.4 percent of common benchmark data sets Adience and IJB-A, and thus benchmarking on them failed to discover and penalize underperformance on this part of the population. Since this study was published, other algorithms have been benchmarked on more balanced face data sets, changing the development process itself to encourage models that perform well across groups.[24]

3.7 Deployment Bias

Deployment bias arises when there is a mismatch between the problem a model is intended to solve and the way in which it is actually used. This often occurs when a system is built and evaluated as if it were fully autonomous, while in reality, it operates in a complicated sociotechnical system moderated by institutional structures and human decision makers. (Andrew Selbst and colleagues refer to this as the "framing trap."[25]) In some cases, for example, systems produce results that must first be interpreted by human decision makers. Despite good performance in isolation, they may end up causing harmful consequences because of phenomena such as automation or confirmation bias.

3.7.1. Example: Risk Assessment Tools in Practice

Algorithmic risk assessment tools in the criminal justice context (also described in Section 3.3.1) are models intended to predict a person's likelihood of committing a future crime. In practice, however, these tools may be used in "off-label" ways, such as to help determine the length of a sentence. Erin Collins describes the harmful consequences of risk assessment tools for actuarial sentencing, including justifying increased incarceration on the basis of personal characteristics.[26] Megan Stevenson builds on this idea, and through an in-depth study of the deployment of risk assessment tools in Kentucky, demonstrates how evaluating the system in isolation created unrealistic notions of its benefits and consequences.[27]

3.8. Identifying Sources of Harm

Knowledge of a model's context and intended use should inform identifying and addressing sources of harm. Recognizing historical bias, for example, requires a retrospective understanding of how structural oppression has manifested in a particular domain over time. Issues that arise in image recognition are frequently related to representation or evaluation bias since many large publicly available image data sets and benchmarks are collected via web scraping, and thus do not equally represent different groups, objects, or geographies. When features or labels represent human decisions (e.g., diagnoses in the medical context, human-assigned ratings in the hiring context), they typically serve as proxies for some underlying, unmeasurable concept, and may introduce measurement bias.

Identifying aggregation bias usually requires some understanding of meaningful underlying groups in the data and reason to think they have different conditional distributions with respect to the prediction label. Medical applications, for example,

often risk aggregation bias because patients of different sexes with similar underlying conditions may present and progress in different ways. Deployment bias is often a concern when systems are used as decision aids for people, since the human intermediary may act on predictions in ways that are typically not modeled in the system.

3.9. Designing Mitigations

Different sources of harm necessitate different interventions. Understanding where intervention is necessary and how feasible it is can inform discussions around when harm can be mitigated versus when it is better not to deploy a system at all. We provide a broad overview of different mitigation approaches here, and a more rigorous discussion in the appendix.

As an example, measurement bias will not be solved by collecting more data, since the underlying issues with the way that data is measured will still be present—but data collection might be useful for combating representation bias. On the other hand, since measurement bias is an issue with the way that features and labels are produced, it may be targeted with more thoughtful and context-aware annotation, as in the work by Desmond Patton and collaborators.[28] It is important to note when there are multiple sources of bias—for example, if both representation and measurement bias are present, simply collecting more data that undergoes the same measurement process would not be sufficient.

Mitigations for other sources of bias might target the model-building process. For example, since aggregation bias is an issue with the model definition, we might combat it by building separate models for different groups in the data, or by parameterizing the model in a way that takes into account underlying differences (such as a multitask model).[29] Learning bias is an issue within the optimization process, and should be targeted by understanding the effect of and experimenting with different objective functions. Evaluation bias stems from underrepresentation in benchmark or test data or the metrics that are used to measure performance. It might be targeted through collecting more diverse, representative benchmarks or through reporting more detailed, granular performance metrics.[30]

Deployment bias is an issue with the real-world scaffolding around a particular system. It can be targeted through changing the model, such as defining a more human-interpretable model that is more easily integrated into the real-world setting. Or it may be targeted through improving the deployment process, such as educating users about

how the model should and should not be used or designing interfaces that facilitate appropriate use of the model.

4. Conclusion

This case study provides a framework for understanding the sources of downstream harm caused by ML systems. We do so in a way that we hope will facilitate productive communication around these issues; 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 what assumptions they are making about the data and domain.

By framing sources of downstream harm through the data generation, model building, evaluation, and deployment processes, 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 broader data generation and ML pipeline that extend far beyond just model training. In practice, ML is an iterative process with a long and complicated feedback loop. We highlight problems that manifest throughout, from historical context to the process of benchmarking models to their final integration into real-world systems.

Discussion Questions and Activities

• Can you come up with an additional example of each source of harm?

• Can you come up with additional ways to detect and/or mitigate each source of harm?

• How might these sources of harm manifest in a project you are currently working on, or have worked on in the past?

• Can you think of other sources of harm that aren’t captured in this case study?

• Draft a checklist that people starting new projects could use to try to anticipate or prevent different types of bias.

• Think about a specific ML-based project or product (something that you’ve worked on, used, or are familiar with) and think about how you might conduct each step of the data collection, development, and deployment process while being conscious of potential sources of harm.

Appendix: Formalizations and Mitigations

In this appendix, we take a step toward formalizing some of the notions introduced in the previous sections. We do this by abstracting the ML pipeline to a series of data transformations. This formalization provides a context we then use to discuss targeted mitigations for specific sources of bias.

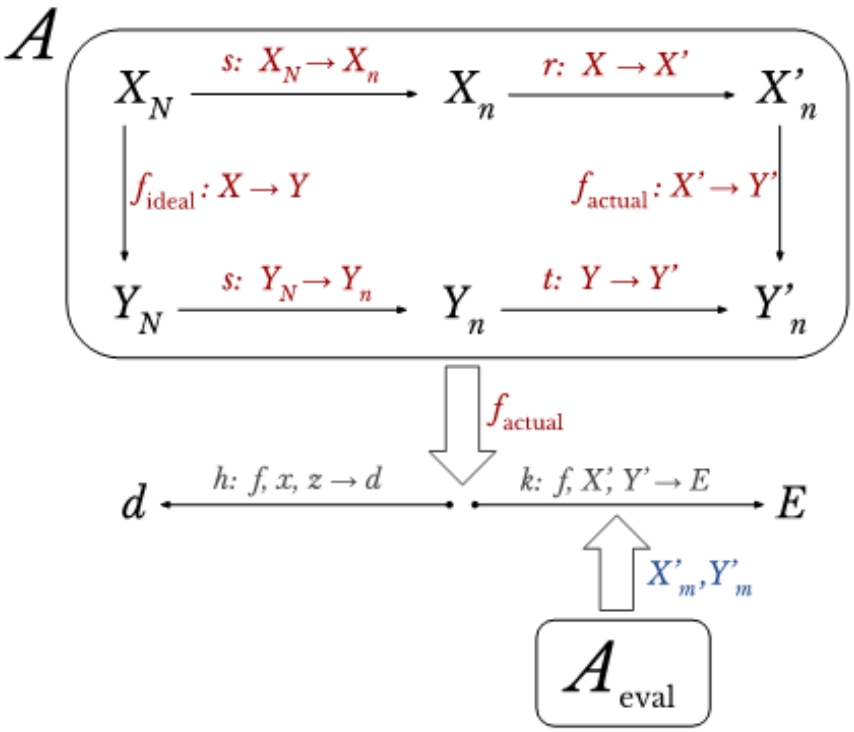


Figure A.1. A data generation and machine learning pipeline viewed as a series of mapping functions. The upper part of the diagram deals with data collection and model building, while the bottom half describes the evaluation and deployment process. See the text for a detailed description.

Consider the data transformations for a data set as depicted in Figure A.1. This data transformation sequence can be abstracted into a general process A. Let X and Y be the underlying feature and label constructs we wish to capture. The subscript indicates the size of the populations, so X_N indicates these constructs over the target population and X_n indicates the smaller development sample, where $s : X_N \rightarrow X_n$ is the sampling function. X' and Y' are the measured feature and label proxies that are chosen to build a model, where r and t are the projections from constructs to proxies, that is, $X \rightarrow X'$ and $Y \rightarrow Y'$. The function $f_{\text{ideal}} : X \rightarrow Y$ is the target function—learned using the ideal

constructs from the target population—but $f_{\text{actual}} : X' \rightarrow Y'$ is the actual function that is learned using proxies measured from the development sample. Then, the function k computes some evaluation metric(s) E for f_{actual} on data X'_m, Y'_m (possibly generated by a different process, e.g., A_{eval} in Figure A.1). Finally, given the learned function f_{actual}, a new input example x, and any external, environmental information z, a function h governs the real-world decision d that will be made (e.g., a human decision maker taking a model's prediction and making a final decision).

There is a growing body of work on "fairness-aware algorithms" that modify some part of the modeling pipeline to satisfy particular notions of "fairness." Interested readers are referred to the review by Arvind Narayanan for a detailed overview of different fairness definitions typically found in this literature, and the work by Sorelle Friedler and collaborators for a comparison of several of these techniques on a number of different data sets.[31] Our aim is to understand classes of particular mitigation techniques in terms of their ability to target different *sources* of problems. In doing so, we can get a better understanding of when and why these approaches might help, and what hidden assumptions they make.

As an example, measurement bias is related to how features and labels are generated (i.e., how r and t are instantiated). Historical bias is defined by inherent problems with the distribution of X and/or Y across the entire population. Therefore, solutions that try to adjust s by collecting more data (that then undergoes the same transformation to X') will likely be ineffective for either of these issues. However, systematically over- or undersampling to change the distributions of X and Y may be able to help with historical bias; in the case of measurement bias, changing r and t through better measurement or annotation processes might be helpful.

In contrast, representation bias stems either from the target population definition (X_N, Y_N) or the sampling function (s). In this case, 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 and that, therefore, improving s will mitigate the harm.

Learning bias is an issue with the way f is optimized, and mitigations should target the defined objective(s) and learning process. In addition, some sources of harm are connected: for example, learning bias can exacerbate performance disparities on

underrepresented groups, so changing s to more equally represent different groups/examples could also help prevent it.

Deployment bias arises when h introduces unexpected behavior affecting the final decision d. Dealing with deployment bias is challenging since the function h is usually determined by complex real-world institutions or human decision makers. Mitigating deployment bias might involve instituting a system of checks and balances in which users balance their faith in model predictions with other information and judgments. This might be facilitated by choosing an f' that is human-interpretable, or by developing interfaces that help users understand model uncertainty and how predictions should be used.

Evaluation and aggregation bias are discussed in more detail in the following.

Example A: Mitigating Aggregation Bias

Aggregation bias is a limitation on the learned function f (e.g., a linear parameterization) that stems from an assumption about the homogeneity of $p(Y' \mid X')$. This results in an f that is disproportionately worse for some group(s). Addressing limitations of f can be achieved by either 1) parameterizing f so that it better models the data complexities, or 2) transforming the training data such that f is now better suited to it.

Methods that adjust f include *coupled learning methods*, such as multitask learning, that parameterize different groups differently in the model definition and facilitate learning multiple functions that take into account group differences.[32]

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 $r: X \rightarrow X'$) in which 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 $p(Y' \mid X')$ is the same across groups), and then learning f.[33] Note that solutions such as *anti-classification* that make predictions independently of group membership do not address aggregation bias.[34]

Example B: Mitigating Evaluation Bias

Evaluation bias is an issue with E, a measurement of the quality of the learned function, f. If we trace the inputs to E, we can see that addressing it would involve 1) redefining k (the function that computes evaluation metrics) and/or 2) adjusting the data X' and Y' on which metrics are computed.

Improving k involves making it more *comprehensive* and *granular*. The granularity of k could be improved with *subgroup evaluation* that compares per-group metrics as well as aggregate measures that weight groups equally.[35] Deciding what groups to use is often application-specific and requires intersectional analysis and privacy considerations.[36]

Multiple metrics and confidence intervals could target 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).[37]

Issues with evaluation data X'_m and Y'_m stem from problems within the data generation process in A_{eval}, for example, an unrepresentative sampling function s_{eval}. Improving s_{eval} could involve *targeted data augmentation* to populate parts of the data distribution that are underrepresented.[38] In other cases, it may be better to develop entirely *new benchmarks* that are more representative and better suited to the task at hand.[39]

Bibliography

Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner. "Machine Bias." *ProPublica,* May 23, 2016.

Atwood, James, Yoni Halpern, Pallavi Baljekar, Eric Breck, D. Sculley, Pavel Ostyakov, Sergey I. Nikolenko, Igor Ivanov, Roman Solovyev, Weimin Wang, and Miha Skalic. "The Inclusive Images Competition." In *The NeurIPS '18 Competition*, edited by S. Escalera and R. Herbrich, 155–86. New York: Springer, 2020.

Bagdasaryan, Eugene, Omid Poursaeed, and Vitaly Shmatikov. "Differential Privacy Has Disparate Impact on Model Accuracy." *Advances in Neural Information Processing Systems* 32 (2019): 15479–88.

Barocas, Solon, Kate Crawford, Aaron Shapiro, and Hanna Wallach. "The Problem with Bias: From Allocative to Representational Harms in Machine Learning." Presented at the 9th Annual Conference of the *Special Interest Group for Computing, Information and Society (SIGCIS)*, Philadelphia, PA, October 29, 2017.

Barocas, Solon, and Andrew D. Selbst. "Big Data's Disparate Impact." *California Law Review* 104 (2016): 671–732.

Buolamwini, Joy, and Timnit Gebru. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification." Proceedings of the 1st Conference on Fairness, Accountability and Transparency. *PMLR* 81 (2018): 77–91.

Calderone, Karen L. "The Influence of Gender on the Frequency of Pain and Sedative Medication Administered to Postoperative Patients." *Sex Roles* 23 (1990): 713–25.

Chawla, Nitesh V., Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. "SMOTE: Synthetic Minority Over-Sampling Technique." *Journal of Artificial Intelligence Research* 16, no. 1 (2002): 321–57.

Chen, Irene Y., Fredrik D. Johansson, and David Sontag. "Why Is My Classifier Discriminatory?" *NIPS'18: Proceedings of the 32nd International Conference on Neural Information Processing Systems* (December 2018): 3543–54.

Collins, Erin. "Punishing Risk." *Georgetown Law Journal* 107 (2018): 57–108.

Corbett-Davies, Sam, Sharad Goel, Jamie Morgenstern, and Rachel Cummings. "Defining and Designing Fair Algorithms." *EC'18: Proceedings of the 2018 ACM Conference on Economics and Computation* (June 2018): 705.

De-Arteaga, Maria, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. "Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Setting." *FAT*'19: Proceedings of the Conference on Fairness, Accountability, and Transparency* (January 2019): 120–28.

Deng, Jia, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. "ImageNet: A Large-Scale Hierarchical Image Database." *2009 IEEE Conference on Computer Vision and Pattern Recognition* (2019): 248–55.

Denton, Emily, Alex Hanna, Razvan Amironesei, Andrew Smart, Hilary Nicole, and Morgan Klaus Scheuerman. "Bringing the People Back in: Contesting Benchmark Machine Learning Datasets," Preprint, submitted July 14, 2020. https://arxiv.org/abs/2007.07399.

Dressel, Julia, and Hany Farid. "The Accuracy, Fairness, and Limits of Predicting Recividism." *Science Advances* 4, no. 1 (2018): eaao5580.

Dressel, Julia, and Hany Farid. "The Dangers of Risk Prediction in the Criminal Justice System." *MIT Case Studies in Social and Ethical Responsibilities of Computing* (February 2021).

Dwork, Cynthia, Nichole Immorlica, Adam Tauman Kalai, and Max Leiserson. "Decoupled Classifiers for Group-Fair and Efficient Machine Learning." Proceedings of the 1st Conference on Fairness, Accountability and Transparency. *PMLR* 81 (2018): 119–33.

Ensign, Danielle, Sorelle A. Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian. "Runaway Feedback Loops in Predictive Policing." *Proceedings of the 1st Conference on Fairness, Accountability and Transparency. PMLR* 81 (2018): 160–71.

Frey, William R., Desmond U. Patton, Michael B. Gaskell, and Kyle A. McGregor. "Artificial Intelligence and Inclusion: Formerly Gang-Involved Youth as Domain Experts for Analyzing Unstructured Twitter Data." *Social Science Computer Review* 38, no. 1 (2020): 42–56.

Friedler, Sorelle A., Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choundary, Evan P. Hamilton, and Derek Roth. "A Comparative Study of Fairness-Enhancing Interventions in Machine Learning." *FAT*'19: Proceedings of the Conference on Fairness, Accountability, and Transparency* (January 2019): 329–38.

Garg, Nikhil, Londa Schiebinger, Dan Jurafsky, and James Zou. "Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes." *Proceedings of the National Academy of Sciences* 115 (2018): E3636–44.

Henry, Matt. 2019. "Risk Assessment: Explained." *The Appeal*, December 14, 2019.

Hoffman, Kelly M., Sophie Trawalter, Jordan R. Axt, and M. Norman Oliver. "Racial Bias in Pain Assessment and Treatment Recommendations, and False Beliefs about Biological Differences between Blacks and Whites." *Proceedings of the National Academy of Sciences* 113, no. 16 (2016): 4296–301.

Hoffmann, Diane E. and Anita J. Tarzian. "The Girl Who Cried Pain: A Bias Against Women in the Treatment of Pain." *Journal of Law, Medicine & Ethics* 29 (2001): 13–27.

Hooker, Sara. "Moving Beyond 'Algorithmic Bias Is a Data Problem'." *Patterns* 2, no. 4 (2021): 100241.

Hooker, Sara, Nyalleng Moorosi, Gregory Clark, Samy Bengio, and Emily Denton. "Characterising Bias in Compressed Models," Preprint, submitted October 6, 2020. https://arxiv.org/abs/2010.03058.

Karkkainen, Kimmo, and Jungseock Joo. "FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation." *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (2021): 1548–58.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan. "Algorithmic Fairness." *AEA Papers and Proceedings* 108 (2018): 22–27.

Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan. "Inherent Trade-Offs in the Fair Determination of Risk Scores." *8th Innovations in Theoretical Computer Science Conference* (2017): article 43.

Li, Min, Amy Mickel, and Stanley Taylor. "'Should This Loan Be Approved or Denied?': A Large Dataset with Class Assignment Guidelines." *Journal of Statistics Education* 26 (2018): 55–66.

Mitchell, Margaret, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. "Model Cards for Model Reporting." *FAT*'19: Proceedings of the Conference on Fairness, Accountability, and Transparency* (January 2019): 220–29.

Mossey, Jana M. "Defining Racial and Ethnic Disparities in Pain Management." *Clinical Orthopaedics and Related Research* 469, no. 7 (2011): 1859–70.

Narayanan, Arvind. "Tutorial: 21 Fairness Definitions and Their Politics." Lecture delivered at the *Conference on Fairness, Accountability, and Transparency*, New York City, February 2018.

Patton, Desmond U., William R. Frey, Kyle A. McGregor, Fei-Tzin Lee, Kathleen McKeown, and Emanuel Moss. "Contextual Analysis of Social Media: The Promise and Challenge of Eliciting Context in Social Media Posts with Natural Language Processing." *AIES'20: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (February 2020): 337–42.

Phelan, Sean M., Diane J. Burgess, Mark W. Yeazel, Wendy L. Hellerstedt, Joan M. Griffin, and Michelle van Ryn. "Impact of Weight Bias and Stigma on Quality of Care and Outcomes for Patients with Obesity." *Obesity Reviews* 16, no. 4 (2015): 319–26.

Raghavan, Manish, Solon Barocas, Jon Kleinberg, and Karen Levy. "Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices." In *FAT*'20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (January 2020): 469–81.

Ryu, Hee Jung, Hartwig Adam, and Margaret Mitchell. "InclusiveFaceNet: Improving Face Attribute Detection with Race and Gender Diversity," Preprint, submitted December 1, 2017. https://arxiv.org/abs/1712.00193.

Salzberg, Steven L. "On Comparing Classifiers: Pitfalls to Avoid and a Recommended Approach." *Data Mining and Knowledge Discovery* 1 (1997): 317–28.

Selbst, Andrew D., Danah Boyd, Sorelle A. Fredler, Suresh Venkatasubramanian, and Janet Vertesi. "Fairness and Abstraction in Sociotechnical Systems." *FAT*'19: Proceedings of the Conference on Fairness, Accountability, and Transparency* (January 2019): 59–68.

Shankar, Shreya, Yani Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D. Sculley. "No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World," Preprint, submitted November 22, 2017. https://arxiv.org/abs/1711.08536.

Stevenson, Megan. "Assessing Risk Assessment in Action." *Minnesota Law Review* 103 (2018): 303–84.

Suresh, Harini, Jen J. Gong, and John V. Guttag. "Learning Tasks for Multitask Learning: Heterogeneous Patient Populations in the ICU." *KDD'18: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (July 2018): 802–10.

Zemel, Rich, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. "Learning Fair Representations." *Proceedings of the 30th International Conference on Machine Learning. PMLR* 28, no. 3 (2013): 325–33.

Footnotes

1. On facial analysis systems, see Joy Buolamwini and Timnit Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," *PMLR* 81 (2018): 77–91; on pretrial risk assessment models, see Julia Angwin et al., "Machine Bias," *ProPublica,* May 23, 2016. ↩

2. https://twitter.com/ylecun/status/1274782757907030016?s=20. ↩
3. Solon Barocas et al., "The Problem with Bias: From Allocative to Representational Harms in Machine Learning," presented at the 9th Annual Conference of the *Special Interest Group for Computing, Information and Society (SIGCIS)*, Philadelphia, PA, October 29, 2017. ↩
4. Maria De-Arteaga *et al.*, "Bias In Bios: A Case Study Of Semantic Representation Bias In A High-Stakes Setting," in *FAT*'19: Proceedings of the Conference on Fairness, Accountability, and Transparency* (January 2019): 120–28; Manish Raghavan et al. "Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices," in *FAT*'20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (January 2020): 469–81. ↩
5. Min Li, Amy Mickel, and Stanley Taylor, "'Should This Loan Be Approved Or Denied?': A Large Dataset with Class Assignment Guidelines," *Journal of Statistics Education* 26 (2018): 55–66. ↩
6. Nikhil Garg et al., "Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes," *Proceedings of the National Academy of Sciences* 115 (2018): E3636–44. ↩
7. Jia Deng *et al.,* "Imagenet: A Large-Scale Hierarchical Image Database," *2009 IEEE Conference on Computer Vision and Pattern Recognition* (2009): 248–55. ↩
8. Shreya Shankar *et al.,* "No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World," preprint, submitted November 22, 2017. https://arxiv.org/abs/1711.08536. ↩
9. Jon Kleinberg *et al.,* "Algorithmic Fairness," *AEA Papers and Proceedings* 108 (May 2018): 22–27. ↩
10. Solon Barocas and Andrew D. Selbst, "Big Data's Disparate Impact," *California Law Review* 104 (2016): 671–732; Danielle Ensign *et al.,* "Runaway Feedback Loops in Predictive Policing," *Proceedings of Machine Learning Research* 81 (2018): 160–71. ↩
11. Jana M. Mossey, "Defining Racial and Ethnic Disparities in Pain Management," *Clinical Orthopaedics and Related Research* 469, no. 7 (2011): 1859–70; Sean M. Phelan *et al.,* "Impact of Weight Bias and Stigma on Quality of Care and Outcomes for Patients with Obesity," *Obesity Reviews* 16, no. 4 (2015): 319–26; Diane E.

Hoffmann and Anita J. Tarzian, "The Girl Who Cried Pain: A Bias against Women in the Treatment of Pain," *Journal of Law, Medicine & Ethics* 29 (2001): 13–27. ↩

12. Karen L. Calderone, "The Influence of Gender on the Frequency of Pain and Sedative Medication Administered to Postoperative Patients," *Sex Roles* 23 (1990): 713–25; Kelly M. Hoffman *et al.,* "Racial Bias in Pain Assessment and Treatment Recommendations, and False Beliefs about Biological Differences between Blacks and Whites," *Proceedings of the National Academy of Sciences* 113, no. 16 (2016): 4296–01. ↩
13. Matt Henry, "Risk Assessment: Explained," *The Appeal,* December 14, 2019. ↩
14. Angwin *et al.,* "Machine Bias," *ProPublica,* May 23, 2016. ↩
15. Julia Dressel and Hany Farid, "The Accuracy, Fairness, and Limits of Predicting Recidivism," *Science Advances* 4, no. 1 (2018): eaao5580; see also Dressel and Farid, "The Dangers of Risk Prediction in the Criminal Justice System," *MIT Case Studies in Social and Ethical Responsibilities of Computing* (February 2021). ↩
16. Desmond U. Patton *et al.,* "Contextual Analysis of Social Media: The Promise and Challenge of Eliciting Context in Social Media Posts with Natural Language Processing," *AIES'20: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (February 2020): 337–42. ↩
17. William R. Frey *et al.,* "Artificial Intelligence and Inclusion: Formerly Gang-Involved Youth as Domain Experts for Analyzing Unstructured Twitter Data," *Social Science Computer Review* 38, no. 1 (2020): 42–56. ↩
18. Sara Hooker, "Moving Beyond 'Algorithmic Bias Is a Data Problem'," *Patterns* 2, no. 4 (2021): 100241. ↩
19. Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan, "Inherent Trade-offs in the Fair Determination of Risk Scores," *8th Innovations in Theoretical Computer Science Conference* (2017): article 43. ↩
20. Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov, "Differential Privacy Has Disparate Impact on Model Accuracy," *Advances in Neural Information Processing Systems* 32 (2019): 15479–88. ↩
21. Sara Hooker *et al.,* "Characterising Bias in Compressed Models," preprint, submitted October 6, 2020. https://arxiv.org/abs/2010.03058. ↩

22. Steven L. Salzberg, "On Comparing Classifiers: Pitfalls to avoid and a Recommended Approach," *Data Mining and Knowledge Discovery* 1 (1997): 317–28. ↩
23. Buolamwini and Gebru, "Gender Shades." ↩
24. Hee Jung Ryu, Hartwig Adam, and Margaret Mitchell, "InclusiveFaceNet: Improving Face Attribute Detection with Race and Gender Diversity," preprint, submitted December 1, 2017. https://arxiv.org/abs/1712.00193. ↩
25. Andrew D. Selbst *et al.,* "Fairness and Abstraction in Sociotechnical Systems," *FAT*'19: Proceedings of the Conference on Fairness, Acountability, and Transparency* (January 2019): 59–68. ↩
26. Erin Collins, "Punishing Risk," *Georgetown Law Journal* 107 (2018): 57–108. ↩
27. Megan Stevenson, "Assessing Risk Assessment in Action," *Minnesota Law Review* 103 (2018): 303–84. ↩
28. Patton *et al.,* "Contextual Analysis of Social Media." ↩
29. Cynthia Dwork *et al., "*Decoupled Classifiers for Group-Fair and Efficient Machine Learning," *Proceedings of Machine Learning Research* 81 (2018): 119–33. ↩
30. On representative benchmarks, see James Atwood *et al.,* "The Inclusive Images Competition," in *The NeurIPS '18 Competition*, ed. S. Escalera and R. Herbrich (New York: Springer, 2020), 155–86; and Kimmo Karkkainen and Joo Jungseock, "FairFace: Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation," *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (2021): 1548–58. On granular performance metrics, see Margaret Mitchell *et al,* "Model Cards for Model Reporting," *FAT*'19: Proceedings of the Conference on Fairness, Accountability, and Transparency* (January 2019): 220–29. ↩
31. Arvind Narayanan, "Tutorial: 21 Fairness Definitions and Their Politics," lecture delivered at the 2018 *Conference on Fairness, Accountability, and Transparency, February,* New York City; Sorelle A. Friedler *et al.,* "A Comparative Study of Fairness-Enhancing Interventions in Machine Learning," *FAT*'19: Proceedings of the Conference on Fairness, Accountability, and Transparency* (January 2019): 329–38. ↩
32. Dwork *et al.,* "Decoupled Classifiers"; Harini Suresh, Jen J. Gong, and John V. Guttag, "Learning Tasks for Multitask Learning: Heterogeneous Patient Populations

in the ICU," *KDD'18: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (July 2018): 802–10. ↩

33. Rich Zemel *et al.,* "Learning Fair Representations," *PMLR* 28, no. 3 (2013): 325–33. ↩

34. On "anti-classification," see Sam Corbett-Davies *et al.,* "Defining and Designing Fair Algorithms," *EC'18: Proceedings of the 2018 ACM Conference on Economics and Computation* (June 2018): 705. ↩

35. Buolamwini and Gebru, "Gender Shades." ↩

36. See Mitchell *et al.,* "Model Cards for Model Reporting," for more in-depth discussion. ↩

37. Section 4.4 of Mitchell *et al.,* "Model Cards for Model Reporting" further discusses different metrics. ↩

38. Nitesh V. Chawla *et al.,* "SMOTE: Synthetic Minority Over-Sampling Technique," *Journal of Artificial Intelligence Research* 16, no. 1 (2002): 321–57; Irene Y. Chen, Fredrick D. Johansson, and David Sontag, "Why Is My Classifier Discriminatory?," *NIPS'18: Proceedings of the 32nd International Conference on Neural Information Processing Systems* (December 2018): 3543–54. ↩

39. Emily Denton *et al.,* "Bringing the People Back in: Contesting Benchmark Machine Learning Datasets," preprint, submitted July 14, 2020. https://arxiv.org/abs/2007.07399; Karkkainen and Joo, "FairFace"; and Atwood *et al.,* "Inclusive Images Competition." ↩