DOI:10.1145/3464903

Article development led by acmqueue
queue.acm.org

A survey for practitioners.

BY RAMYA SRINIVASAN AND AJAY CHANDER

Biases in AI Systems

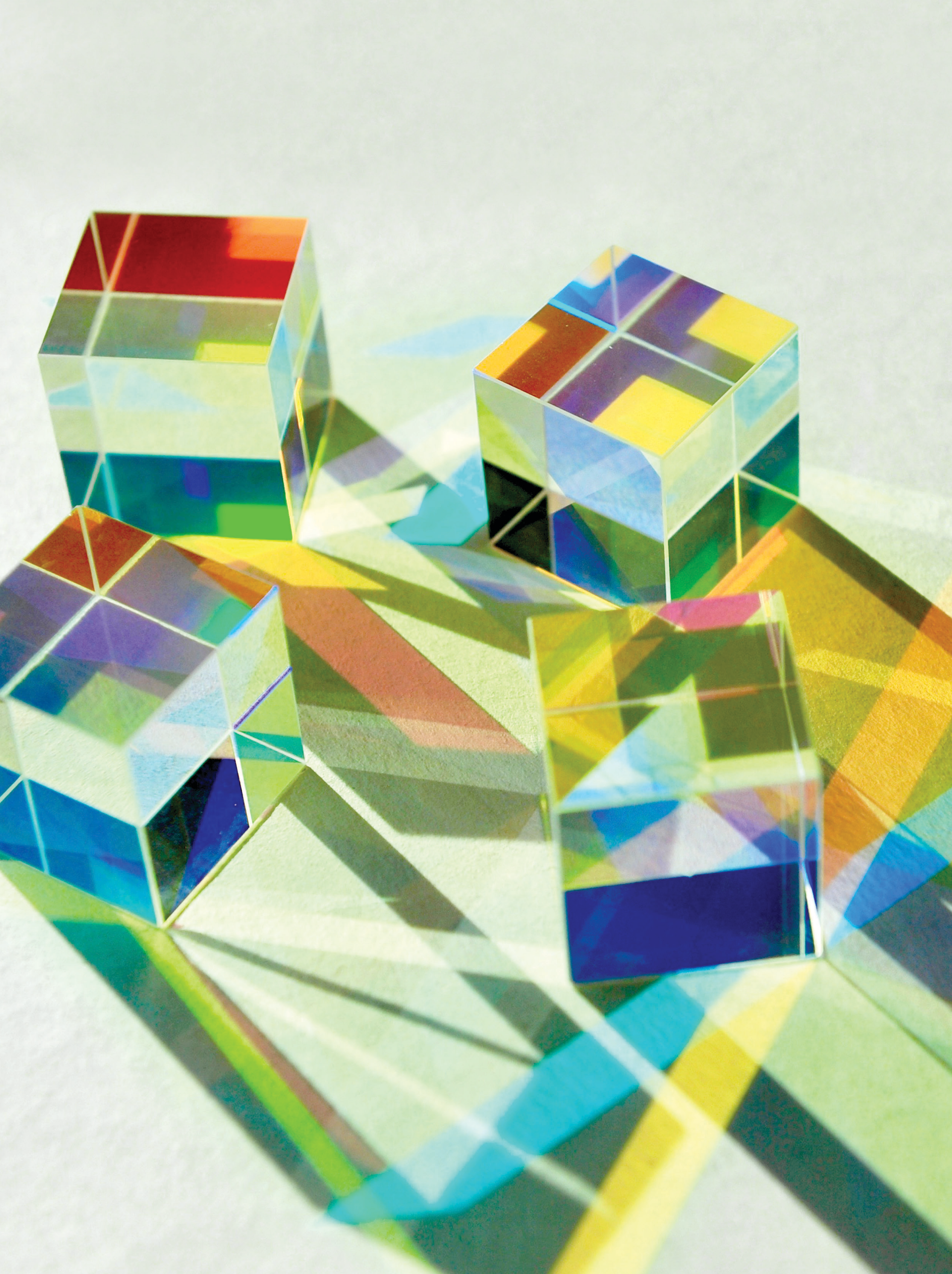
A CHILD WEARING sunglasses is labeled as a "failure, loser, nonstarter, unsuccessful person." This is just one of the many systemic biases exposed by ImageNet Roulette, an art project that applies labels to user-submitted photos by sourcing its identification system from the original ImageNet database.[7] ImageNet, which has been one of the instrumental datasets for advancing AI, has deleted more than half a million images from its "person" category since this instance was reported in late 2019.[23] Earlier in 2019, researchers showed how Facebook's ad-serving algorithm for deciding who is shown a given ad exhibits discrimination based on race, gender, and religion of users.[1] There have been reports of commercial facial-recognition software (notably Amazon's Rekognition, among others) being biased against darker-skinned women.[6,22]

These examples provide a glimpse into a rapidly growing body of work that is exposing the bias associated with AI systems, but biased algorithmic systems are not a new phenomenon. As just one example, in 1988, the U.K. Commission for Racial Equality found a British medical school guilty of discrimination because the algorithm used to shortlist interview candidates was biased against women and applicants with non-European names.[17]

With the rapid adoption of AI across a variety of sectors, including in areas such as justice and health care, technologists and policy makers have raised concerns about the lack of accountability and bias associated with AI-based decisions. From AI researchers and software engineers to product leaders and consumers, a variety of stakeholders are involved in the AI pipeline. The necessary expertise around AI, datasets, and the policy and rights landscape that collectively helps uncover bias is not uniformly available among these stakeholders. As a consequence, bias in AI systems can compound inconspicuously.

Consider, for example, the critical role of machine learning (ML) developers in this pipeline. They are asked to: preprocess the data appropriately, choose the right models from several available ones, tune parameters, and adapt model architectures to suit the requirements of an application. Suppose an ML developer was entrusted with developing an AI model to predict which loans will default. Unaware of bias in the training data, an engineer may inadvertently train models using only the validation accuracy. Suppose the training data contained too many young people who defaulted. In this case, the model is likely to make a similar prediction about young people defaulting when applied to test data. There is thus a need to educate ML developers about

IMAGE BY SUPERPUAY/SHUTTERSTOCK



the various kinds of biases that can creep into the AI pipeline.

Defining, detecting, measuring, and mitigating bias in AI systems is not an easy task and is an active area of research.[4] A number of efforts are being undertaken across governments, nonprofits, and industries, including enforcing regulations to address issues related to bias. As work proceeds toward recognizing and addressing bias in a variety of societal institutions and pathways, there is a growing and persistent effort to ensure that computational systems are designed to address these concerns.

The broad goal of this article is to educate nondomain experts and practitioners such as ML developers about various types of biases that can occur across the different stages of the AI pipeline and suggest checklists for mitigating bias. There is a vast body of literature related to the design of fair algorithms.[4] As this article is directed at aiding ML developers, the focus is not on the design of fair AI algorithms but rather on practical aspects that can be followed to limit and test for bias during problem formulation, data creation, data analysis, and evaluation. Specifically, the contributions can be summarized as follows:

▸ *Taxonomy of biases in the AI pipeline.* A structural organization of the various types of bias that can creep into the AI pipeline is provided, anchored in the various phases from data creation and problem formulation to data preparation and analysis.

▸ *Guidelines for bridging the gap between research and practice.* Analyses that elucidate the challenges associated with implementing research ideas in the real world are listed, as well as suggested practices to fill this gap. Guidelines that can aid ML developers in testing for various kinds of biases are provided.

The goal of this work is to enhance awareness and practical skills around bias, toward the judicious use and adoption of AI systems.

Biases in the AI Pipeline

A typical AI pipeline starts from the data-creation stage: collecting the data; annotating or labeling it; and preparing or processing it into a format that can be consumed by the rest of the pipeline. Let's analyze how different types of bias can be introduced in each of these steps.

Data-creation bias. Specific types of biases can occur during the creation of datasets.

Sampling Bias

The bias that arises in a dataset that is created by selecting particular types of instances more than others (and thereby rendering the dataset underrepresentative of the real world) is called *sampling bias*. This is one of the most common types of dataset biases. Datasets are often created with a particular set of instances. For example, image datasets prefer street scenes or nature scenes.[25] A face-recognition algorithm may be fed with more photos of light-skinned faces than dark-skinned faces, thereby leading to poor performance in recognizing darker-skinned faces.[6] Thus, sampling bias can result in poor generalization of learned algorithms.

Measurement Bias

Measurement bias is introduced by errors in human measurement, or because of certain intrinsic habits of people in capturing data. As an example, consider the creation of image and video datasets, where the images or videos may reflect the techniques used by the photographers. For example, some photographers might tend to take pictures of objects in similar ways; as a result, the dataset may contain object views from certain angles only. In their 2011 paper "Unbiased Look at Dataset Bias," Torralba and Efros refer to this type of measurement bias as *capture bias*.[25]

Another source of measurement bias could be a result of the device used to capture datasets. For example, cameras used to capture images may be defective, leading to poor-quality images and thereby contributing to biased results. These types of biases are broadly categorized as *device bias*.

A third type of measurement bias can occur when proxies are used instead of true values in creating the dataset. For example, arrest rates are often used instead of crime rates; doctor visits and medications are used as indicators of medical conditions, and so on.

Label Bias

Label bias is associated with inconsistencies in the labeling process. Different annotators have different styles and preferences that get reflected in the labels created. A common instance of label bias arises when different annotators assign differing labels to the same type of object (for example, *grass* vs. *lawn*, *painting* vs. *picture*).[25]

Yet another type of label bias can happen when the subjective biases of evaluators affect labeling. For example, in a task of annotating emotions experienced in a text, the labels could be biased by the subjective preferences of annotators such as their culture, beliefs, and introspective capabilities.[24] *Confirmation bias*,[21] which is the human tendency to search for, interpret, focus on, and remember information in a way that confirms one's preconceptions, is closely related to this type of label bias. Thus, labels may be assigned based on prior belief rather than objective assessments.

A third type of label bias can arise from the peak end effect. This is a type of memory-related cognitive bias in which people judge an experience based largely on how they felt at its peak (that is, its most intense point) and at its end, rather than based on the total sum or average of every moment of the experience.[15] For example, some annotators may give more importance to the last part of a conversation (rather than the entire conversation) in assigning a label.[24]

Negative Set Bias

Torralba and Efros define *negative set bias* as being introduced in the dataset as a consequence of not having enough samples representative of "the rest of the world."[25] The authors state that "datasets define a phenomenon (for example, object, scene, event) not just by what it is (positive instances), but also by what it is not (negative instances)." As a consequence, the learned classifiers can perform poorly in detecting negative instances.

Biases related to problem formulation. Biases can arise based on how a problem is defined. Consider the following example presented in *MIT Technology Review* by Karen Hao.[13] Suppose a credit card company wants to predict a customer's creditworthiness using

AI. In order to do so, creditworthiness must be defined in a manner that can be "predicted or estimated." The problem can be formulated based on what the company wants, say, to maximize its profit margin or to maximize the number of loans that get repaid; however, "those decisions are made for various business reasons other than fairness or discrimination," says Cornell University's Solan Barocas, who specializes in fairness.

Framing Effect Bias

The previous creditworthiness example can be thought of as a type of *framing effect bias*.[21] Based on how the problem is formulated and how information is presented, the results obtained can be different and perhaps biased. Another notable example is the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) debate[8] concerning the definition of fairness between Northpointe (now known as Equivant), which came up with COMPAS scores for assessing risk of recidivism, and *ProPublica*, which claimed that the COMPAS system was biased. *ProPublica* claimed that Northpointe's method was biased against black defendants as the group was associated with a higher false-positive rate. There are several metrics of fairness, and *ProPublica* stated that Northpointe's system violated equalized odds and equality of opportunity fairness criteria. Northpointe's main defense was that scores satisfied fairness from the viewpoint of predictive rate parity.[4] Thus, bias can arise based on the way a problem and its success metrics are defined.

Biases related to the algorithm/data analysis. Several types of biases can occur in the algorithm or during data analysis.

Sample Selection Bias

Sample selection bias is introduced by the selection of individuals, groups, or data for analysis in such a way that the samples are not representative of the population intended to be analyzed.[9] In particular, sample selection bias occurs during data analysis as a result of conditioning on some variables in the dataset (for example, a particular skin color, gender, among others), which in turn can create spurious correlations. For example, in analyzing the effect of motherhood on wages, if the study is restricted to women who are already employed, then the measured effect will be biased as a result of conditioning on employed women.[9] Common types of sample selection bias include Berkson's paradox[20] and sample truncation.[9]

Based on how the problem is formulated and how information is presented, the results obtained can be different and perhaps biased.

Confounding Bias

Bias can arise in the AI model if the algorithm learns the wrong relations by not taking into account all the information in th e data or if it misses the relevant relations between features and target outputs.[20] *Confounding bias* originates from common causes that affect both inputs and outputs. Consider a scenario wherein admissions to a graduate school are based on the person's previous grade point average. There might be other factors, however, such as ability to get coaching, which in turn may be dependent on sensitive attributes such as race; and these factors may determine the grade point average and admission rates.[16] As a result, spurious relations between inputs and outputs are introduced and thus can lead to bias.

A special type of confounding bias is the *omitted variable*, which occurs when some relevant features are not included in the analysis. This is also related to the problem of model underfitting.

Another type of confounding bias is the *proxy variable*. Even if sensitive variables such as race and gender are not considered for decision making, certain other variables used in the analysis might serve as "proxies" for those sensitive variables. For example, zip code might be indicative of race, as people of a certain race might predominantly live in a certain neighborhood. This type of bias is also commonly referred to as *indirect bias* or *indirect discrimination*.

Design-Related Bias

Sometimes, biases occur as a result of algorithmic limitations or other constraints on the system such as computational power. A notable entry in this category is *algorithm bias*, which can be defined as bais that is solely induced or added by the algorithm. In their 1996 paper "Bias in Computer Systems," Friedman and Nissenbaum[10] provide an example: Software that relies on randomness for fair distributions of results is not truly random; for example, by skewing selections toward items at the end or beginning of a list, the results can become biased.

Another type of design-related bias is *ranking bias*.[18] For example, a search engine that shows three results per screen can be understood to privilege the top three results slightly more than the next three.[10] Ranking bias is also closely related to presentation bias,[18] which is derived from the fact that you can receive user feedback only on items that have been presented to the user. Even among those that are shown, the probability of receiving user feedback is further affected by where the item is shown.[2]

Biases related to evaluation/validation. Several types of biases result from those inherent in human evaluators, as well as in the selection of those evaluators (sample treatment bias).

Human Evaluation Biases

Often, human evaluators are employed in validating the performance of an AI model. Phenomena such as confirmation bias, peak end effect, and prior beliefs (for example, culture) can create biases in evaluation.[15] Human evaluators are also constrained by how much information they can recall, which can result in *recall bias*.

Sample Treatment Bias

Sometimes, test sets selected for evaluating an algorithm may be biased.[3] For example, in recommendation systems, some specific viewers (for example, those speaking a certain language) may be shown an advertisement, and some may not. As a consequence, the observed effects will not be representative of true effects on the general population. The bias introduced in the process of selectively subjecting some sets of people to a type of treatment is called *sample treatment bias*.

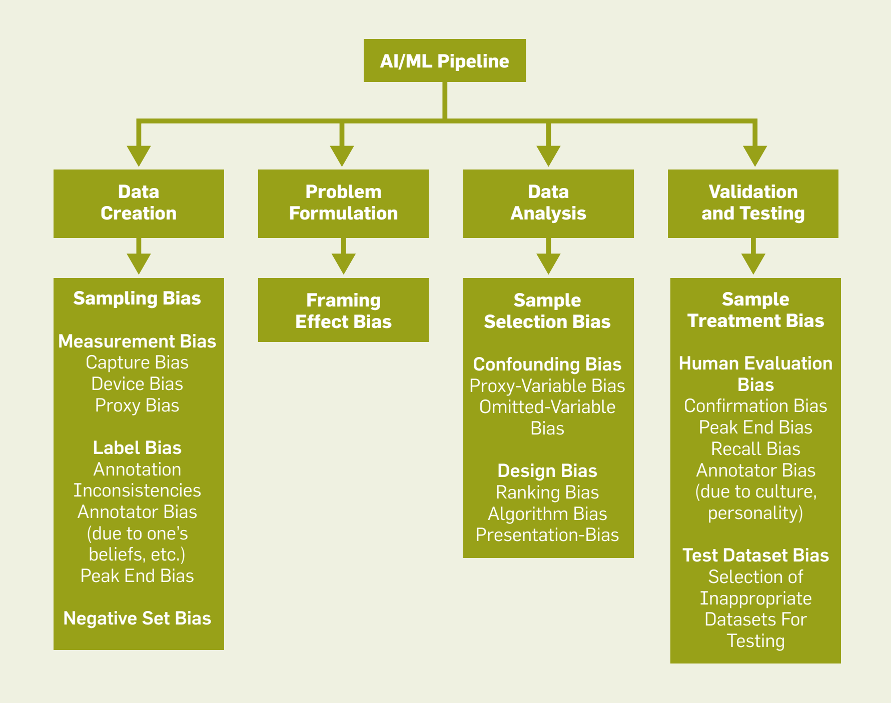
Validation and Test Dataset Biases

Biases can also be induced from sample selection and label biases in the validation and test datasets.[25] In general, biases associated with the dataset-creation stage could show up in the model-evaluation stage as well. Additionally, evaluation bias can result from the selection of inappropriate benchmarks/datasets for testing.

The accompanying figure provides an illustration of the taxonomy of biases along the various stages of the AI pipeline as discussed in the previous sections.

Despite significant research efforts within the AI community to address bias-related challenges, several gaps impede the collective progress. Next, we highlight some of these gaps.

Taxonomy of bias types along the AI pipleline.



Gaps Between Research and Practice

Methods to counter dataset bias issues have been proposed, as have new datasets with an emphasis on maintaining diversity. For example, the diversity-in-faces dataset consists of almost a million images of people pulled from the Yahoo! Flickr Creative Commons dataset, assembled specifically to achieve statistical parity among categories of skin tone, facial structure, age, and gender. In their 2019 paper, "Excavating AI," Crawford and Paglen, however, question the use of cranio-metrical features used in creating this dataset, as these features could also be proxies for racial bias.[7] The authors further provide a critical review of issues pertaining to several benchmark datasets.

"Fairness in machine learning" is an active area of research. There are also conferences and workshops dedicated to the theme. A complete overview of fairness in machine learning is beyond the scope of this survey. For an extensive overview of various algorithmic definitions of fairness and methods to achieve fairness in classification, consult Barocas et al.[4] There are also open-source tools such as IBM's AI Fairness 3605 that facilitates detection and mitigation of unwanted algorithmic bias. Despite these efforts, there are notable gaps, as noted by Gajane and Pechenizkiy in their 2018 paper, "On Formalizing Fairness in Prediction with Machine Learning.[11]

Filling the gap. Practice guidelines have been proposed for reducing the potential bias in AI systems. These include "Factsheets for Datasets" from IBM, and "Datasheets for Datasets," an approach for sharing essential information about datasets used to train AI models.[12] In their 2019 paper, Mitchell et al. suggest the use of detailed documentation of released models in order to encourage transparency.[19]

Holstein et al. identify areas of alignment and disconnect between the challenges faced by teams in practice and the solutions proposed in the fair ML research literature.[14] The authors urge that future research should focus on supporting practitioners in collecting and curating high-quality datasets. The authors further see a need for creating domain-specific educational resources, metrics, processes, and tools. In that spirit, this article aims to be an educational resource for ML developers in understanding various sources of biases in the AI pipeline.

Guidelines for ML Developers

While it may not be possible to eliminate all sources of bias, with certain precautionary measures, some bias issues can be reduced. Here are some key messages that could aid ML developers in identifying potential sources of bias and help in avoiding the introduction of unwanted bias:

▸ Incorporation of domain-specific knowledge is crucial in defining and detecting bias. It is important to understand the structural dependencies among various features in the dataset. Often, it helps to draw a structural diagram illustrating various features of interest and their interdependencies. This can then help in identifying the sources of bias.[20]

▸ It is also important to understand which features of the data are deemed sensitive based on the application. For example, age may be a sensitive feature in determining who gets a loan, but not necessarily in determining who gets a medical treatment. Furthermore, there may be proxy features that, although not thought to be sensitive features, may still encode sensitive information so as to render biased predictions.

▸ As far as possible, datasets used for analysis should be representative of the true population under consideration. Thus, care has to be taken in constructing representative datasets.

▸ Appropriate standards have to be laid out for annotating the data. Rules have to defined so as to obtain consistent labels from annotators as much as possible.

▸ Identifying all features that may be associated with the target feature of interest is important. Omitting variables that have dependencies with the target feature leads to a biased estimate.

▸ Features that are associated with both input and output can lead to biased estimates. In such cases, it is important to eliminate these sources of confounding biases by appropriate data conditioning and randomization strategies in selecting input.[20]

▸ Restricting data analysis to some truncated portions of the dataset can lead to unwanted selection bias. Thus, in choosing subsets of data for analysis, care must be taken not to introduce sample selection bias.

▸ In validating the performance of a model such as in A/B testing, care has to be taken to guard against the introduction of sample treatment bias. In other words, in testing the performance of a model, the test conditions should not be restricted to a certain subset of the population (for example, showing recommendation results to people of a certain locality only), as the results would be biased.

Conclusion

This article provides an organization of various kinds of biases that can occur in the AI pipeline starting from dataset creation and problem formulation to data analysis and evaluation. It highlights the challenges associated with the design of bias-mitigation strategies, and it outlines some best practices suggested by researchers. Finally, a set of guidelines is presented that could aid ML developers in identifying potential sources of bias, as well as avoiding the introduction of unwanted biases. The work is meant to serve as an educational resource for ML developers in handling and addressing issues related to bias in AI systems.

References

1. Ali, M., Sapiezynsk, P., Bogen, M., Korolova, A., Mislove, A., Rieke, A. Discrimination through optimization: how Facebook's ad delivery can lead to biased outcomes. In *Proceedings of the ACM on Human-Computer Interaction 3* (2019); https://dl.acm.org/doi/10.1145/3359301.
2. Amatriain, X. What does the concept of presentation-feedback bias refer to in the context of machine learning? Quora, 2015; https://www.quora.com/What-does-the-concept-of-presentation-feedback-bias-refer-to-in-the-context-of-machine-learning.
3. Austin, P.C., Platt, R.W. Survivor treatment bias, treatment selection bias, and propensity scores in observational research. , 2 (2010), 136–138; https://www.jclinepi.com/article/S0895-4356(09)00247-9/fulltext.
4. Barocas, S., Hardt, M., Narayanan, A. Fairness and machine learning: limitations and opportunities, 2019; https://fairmlbook.org.
5. Bellamy, R.K.E. et al. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. 2018, arXiv; https://arxiv.org/abs/1810.01943.
6. Buolamwini, J., Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of Machine Learning Research 81* (2018), 1–15; http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf.
7. Crawford, K., Paglen, T. Excavating AI: The politics of images in machine learning training sets. The AI Now Institute, New York University, 2019; https://www.excavating.ai.
8. Dressel, J., Farid, H. The accuracy, fairness, and limits of predicting recidivism. *Science Advances 4*, 1 (2018); https://advances.sciencemag.org/content/4/1/eaao5580.
9. Elwert, F., Winship, C. Endogenous selection bias: The problem of conditioning on a collider variable. *Annual Review of Sociology 40* (2014), 31-53; https://www.annualreviews.org/doi/full/10.1146/annurev-soc-071913-043455.
10. Friedman, B., Nissenbaum, H. Bias in computer systems. In *ACM Trans. Information Systems 14*, 3 (1996), https://dl.acm.org/doi/10.1145/230538.230561.
11. Gajane, P., Pechenizkiy, M. On formalizing fairness in prediction with machine learning. In *Proceedings of the Intern. Conf. Machine Learning, Fairness Accountability and Transparency Workshop*, 2018; https://www.fatml.org/media/documents/formalizing_fairness_in_prediction_with_ml.pdf.
12. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., Crawford, K. Datasheets for datasets. In *Proceedings of the 5th Workshop on Fairness, Accountability, and Transparency in Machine Learning*, 2018; https://www.microsoft.com/en-us/research/uploads/prod/2019/01/1803.09010.pdf.
13. Hao, K. This is how AI bias really happens—and why it's so hard to fix. *MIT Technology Review*; https://www.technologyreview.com/2019/02/04/137602/this-is-how-ai-bias-really-happensand-why-its-so-hard-to-fix/.
14. Holstein, K., Vaughan, J.W., Daumé III, H., Dudik, M., Wallach, H. Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of the 2019 SIGCHI Con. Human Factors in Computing Systems*, 1–16; https://dl.acm.org/doi/10.1145/3290605.3300830.
15. Kahneman, D. Evaluation by moments: past and future. *Choices, Values and Frames.* D. Kahneman and A. Tversky, Eds. Cambridge University Press, New York, 2000.
16. Kilbertus, N., Ball, P.J., Kusner, M.J., Weller, A., Silva, R. The sensitivity of counterfactual fairness to unmeasured confounding. In *Proceedings of the 2019 Conf. Uncertainty in Artificial Intelligence*; http://auai.org/uai2019/proceedings/papers/213.pdf.
17. Lowry, S., Macpherson, G. 1988. A blot on the profession. *British Medical J.* Clinical Research Ed. 296, 6623 (1988), 657; https://www.bmj.com/content/296/6623/657.
18. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A. A survey on bias and fairness in machine learning. 2019, arXiv; https://arxiv.org/abs/1908.09635.
19. Mitchell, M. et al. Model cards for model reporting. In *Proceedings of the 2019 AAAI/ACM Conf. AI, Ethics, and Society*; arXiv; https://arxiv.org/abs/1810.03993.
20. Pearl, J., Mackenzie, D. *The Book of Why: The New Science of Cause and Effect.* Basic Books, 2018.
21. Plous, S. *The Psychology of Judgment and Decision Making.* McGraw-Hill, 1993.
22. Raji, I., Buolamwini, J. Actionable auditing: investigating the impact of publicly naming biased performance results of commercial AI products. In *Proceedings of the 2019 AAAI/ACM Conf. AI, Ethics, and Society*, 429–435; https://dl.acm.org/doi/10.1145/3306618.3314244.
23. Small, Z. 600,000 images removed from AI database after art project exposes racist bias. *Hyperallergic*, 2019; https://hyperallergic.com/518822/600000-imagesremoved- from-ai-database-after-art-project-exposesracist- bias/.
24. Srinivasan, R., Chander, A. Crowdsourcing in the absence of ground truth—a case study. In *Proceedings of the 2019 Intern. Conf. Machine Learning Workshop on Human in the Loop Learning*; https://arxiv.org/abs/1906.07254.
25. Torralba, A., Efros, A.A. Unbiased look at dataset bias. In *Proceedings of the 2011 IEEE Con. Computer Vision and Pattern Recognition*, 1521–1528; https://ieeexplore.ieee.org/document/5995347.

Ramya Srinivasan is an AI researcher with Fujitsu Research of America. Her background is in the areas of computer vision, machine learning, explainable AI, and AI ethics.

Ajay Chander leads R&D teams in imagining and building new human-centered technologies and products. His work has spanned transparent AI, AI life assistants, digital healthcare and wellness, software tools design, security, and computational behavior design. He has received ACM's Most Influential Paper of the Decade award.