Patterns

CellPress
OPEN ACCESS

Opinion

Moving beyond "algorithmic bias is a data problem"

Sara Hooker[1,*]
[1]Google Brain, Mountain View, CA, USA
*Correspondence: shooker@google.com
https://doi.org/10.1016/j.patter.2021.100241

A surprisingly sticky belief is that a machine learning model merely *reflects* existing algorithmic bias in the dataset and does not itself contribute to harm. Why, despite clear evidence to the contrary, does the myth of the impartial model still hold allure for so many within our research community? Algorithms are not impartial, and some design choices are better than others. Recognizing how model design impacts harm opens up new mitigation techniques that are less burdensome than comprehensive data collection.

Moving beyond "algorithmic bias is a data problem"

In the absence of intentional interventions, a trained machine learning model *can and does* amplify undesirable biases in the training data. A rich body of work to date has examined these forms of problematic algorithmic bias, finding disparities—relating to race, gender, geo-diversity, and more—in the performance of machine learning models.[1]

However, a surprisingly prevalent belief is that a machine learning model merely *reflects* existing algorithmic bias in the dataset and does not itself contribute to harm. Here, we start out with a deceptively simple question: how does model design contribute to algorithmic bias?

A more nuanced understanding of what contributes to algorithmic bias matters because it also dictates where we spend effort mitigating harm. If algorithmic bias is merely a data problem, the often-touted solution is to de-bias the data pipeline. However, data "fixes" such as re-sampling or re-weighting the training distribution are costly and hinge on (1) knowing *a priori* what sensitive features are responsible for the undesirable bias and (2) having comprehensive labels for protected attributes and *all* proxy variables.

For real-world datasets, satisfying both (1) and (2) is more often than not infeasible. For domains such as images, language, and video, the high dimensionality of the problem and large size of modern datasets make it hard to guarantee all features are comprehensively labeled. Even if we are able to label sensitive attributes at scale such as gender and race, algorithms can still leverage proxy variables to reconstruct the forbidden label. Data collection of even a limited number of protected attributes can be onerous. For example, it is hard to align on a standard taxonomy—categories attributed to race or gender are frequently encoded in inconsistent ways across datasets.[2] Furthermore, procuring labels for these legally protected attributes is often perceived as intrusive leading to noisy or incomplete labels.[3,4]

If we cannot guarantee we have fully addressed bias in data pipeline, the overall harm in a system is a product of the interactions between the data and our model design choices. Here, acknowledging the impact of model design bias can play an important role in curbing harm. Algorithms are not impartial, and some design choices are better than others. Recognizing how model design impacts harm opens up new mitigation techniques that are far less burdensome than comprehensive data collection.

The impact of our model design choices

If you replace algorithmic bias with test-set accuracy, it becomes a much more acceptable stance that our modeling choices—architecture, loss function, optimizer, hyper-parameters—express a preference for final model behavior. Most students of machine learning are familiar with some variation of Figure 1, where varying the degree of a polynomial function leads to trained functions with differing levels of overfitting to the training data.

We are well-versed in the connection between function choice and test-set accuracy because objective functions such as cross-entropy or mean squared error reflect our preference to optimize for high test-set accuracy. Standard loss functions do not explicitly encode preferences for other objectives we care about such as algorithmic bias, robustness, compactness, or privacy. However, just because these desiderata are not reflected does not mean they have ceased to exist. Turing award winner Donald Knuth said that computers "do exactly what they are told, no more and no less." A model can fulfill an objective in many ways, while still violating the spirit of said objective.

Model design choices made to maximize test-set accuracy do not hold static other properties we care about such as robustness and fairness. On the contrary, training a parametric model is akin to having a fixed amount of materials to build a house with. If we decide to use more bricks building a bigger living room, we force the redistribution of the number of bricks available for all other rooms. In the same vein, when we prioritize one objective, whether that be test-set accuracy or additional criteria such as compactness and privacy, we inevitably introduce new trade-offs.

A key reason why model design choices amplify algorithmic bias is because notions of fairness often coincide with how underrepresented protected features are treated by the model. Buolamwini and Gebru[5] find that facial-analysis datasets reflect a preponderance of lighter-skinned subjects, with far higher model error rates for dark-skinned women. Shankar et al.[6] show that models trained on datasets with limited geo-diversity show sharp degradation on data drawn from other locales. Word

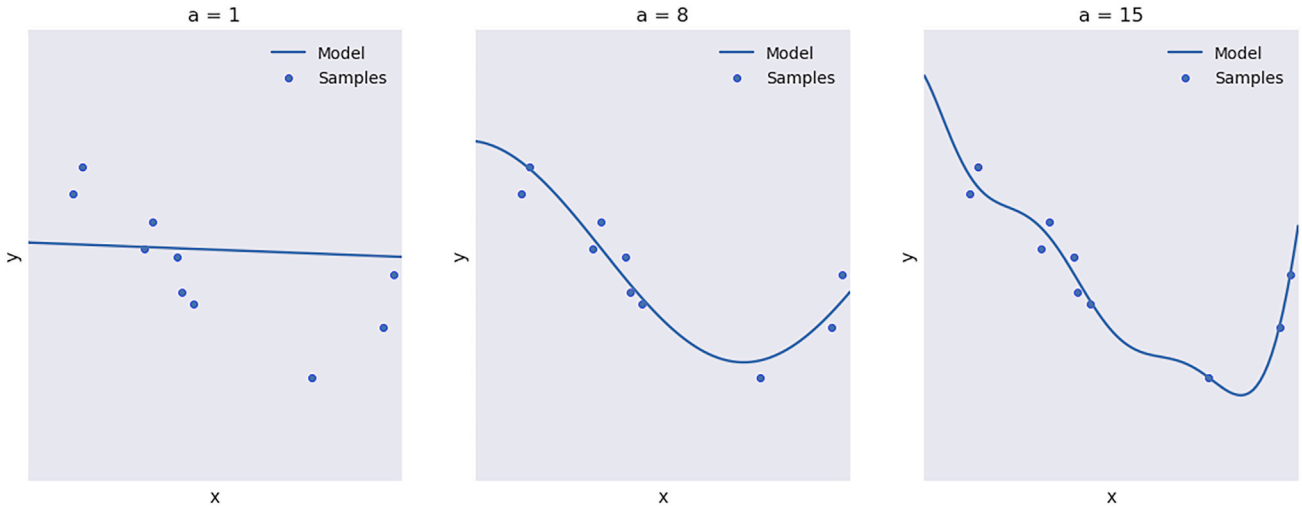

Figure 1. Our model choices express a preference for model behavior. An example most students of machine learning will recognize is the plot between the degrees of a polynomial (a) and the degree of overfitting.

frequency co-occurrences within text datasets frequently reflect social biases relating to gender, race, and disability.[7]

In all these cases, the algorithmic bias a model learns can be attributed to the relative over-and-under representation of a protected attribute within a dataset category. Most real-world data naturally have a skewed distribution similar to the one visualized in Figure 2, with a small number of well-represented features and a "long-tail" of features that are relatively underrepresented. The skew in feature frequency leads to disparate error rates on the underrepresented attribute. This prompts fairness concerns when the underrepresented attribute is a protected attribute but more broadly relates to the brittleness of deep neural network performance in data-limited regimes. Understanding which model design choices disproportionately amplify error rates on protected underrepresented features is a crucial first step in helping curb algorithmic harm.

Measuring complex trade-offs

In complex systems, it is challenging to manipulate one variable in isolation and foresee all implications. Early televised drug prevention advertisements in the 2000s led to increased drug use.[8] The extermination of dogs and cats during the Black Death inadvertently helped spread the disease by accelerating the growth of rat populations.[9] The belief that model design merely *reflects* algorithmic bias in the dataset can be partly ascribed to the difficulty of measuring interactions between all the variables we care about.

This is changing. There is new urgency to scholarship that considers the interactions between multiple model desiderata. Recent work has proposed rigorous frameworks to understand and measure the impact of trade-offs on algorithmic bias. For example: How does optimizing for compactness impact robustness and fairness? What about the trade-off between privacy and fairness?

Recent work has shown that design choices to optimize for either privacy guarantees or compression amplify the disparate impact between minority and majority data subgroups such that the "rich get richer and the poor get poorer." Bagdasaryan et al.[10] show that differential privacy techniques such as gradient clipping and noise injection disproportionately degrade accuracy for darker-skinned faces in the Diversity in Faces (DiF) dataset and users writing tweets in African-American English. My own work with colleagues measures the impact of popular compression techniques like quantization and pruning on low-frequency protected attributes such as gender and age and finds that these subgroups are systematically and disproportionately impacted in order to preserve performance on the most frequent features.[11,12]

These are not the only design choices that matter—even more subtle choices like learning rate and length of training can also disproportionately impact error rates on the long-tail of the dataset. Work on memorization properties of deep neural networks shows that challenging and underrepresented features are learnt later in the training process and that the learning rate impacts what is learnt.[13] Thus, early stopping and similar hyper-parameter choices disproportionately and systematically impact a subset of the data distribution.

A key takeaway is that our algorithms are not impartial. Some design choices are better than others. Given the widespread use of compression and differential privacy techniques in sensitive domains like health care diagnostics, understanding the distribution of error is of paramount importance for auditing potentially adverse harm to human welfare. Here, the trade-offs incurred by pruning or gradient clipping may be intolerable given the impact on human welfare. While these results suggest caution should be used before using these techniques in sensitive domains, it also provides a valuable roadmap to mitigate harm.

For example, a formidable hurdle given the large size of modern training sets is even knowing what to look at to audit for problematic biases. Reasoning about model behavior is often easier when presented with a subset of data points that are more challenging for the model to classify. We can leverage our knowledge about how model design choices exacerbate harm to surface parts of the distribution most likely to require

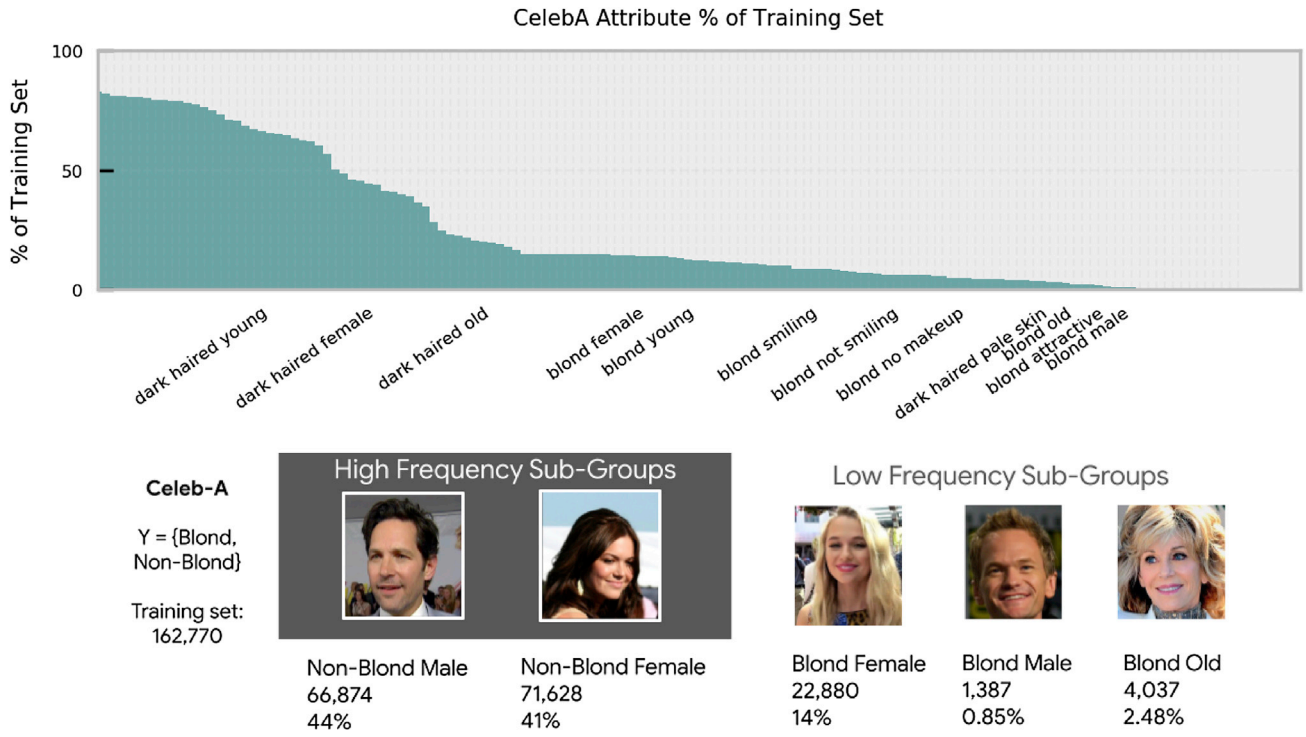

Figure 2. Most natural image datasets exhibit a long-tail distribution with an unequal frequency of attributes in the training data. Notions of fairness often coincide with how underrepresented sensitive attributes are treated by the model. Our model design choices can excacerbate or curb disparate harm on the long-tail.

human auditing. Compression identified exemplars (CIEs) are an example of this human-in-the-loop tooling, surfacing the data points disproportionately impacted by compression. These examples are a small subset of the overall distribution and are identified by comparing predictive behavior of a compressed and non-compressed model (so do not require pre-existing labels for all the features). Inspecting these examples directs limited human auditing time to the most challenging examples and, as shown by recent work by Joseph et al.,[14] can also be used to directly optimize for a model that is both compact *and* less harmful.

Why a more nuanced discussion of the origins of bias matters

Diffusion of responsibility is a socio-psychological phenomenon where an individual abstains from taking action due to the belief that someone else is responsible for intervening. In computer science, diffusion of responsibility often revolves around discussion of what is and isn't "out of scope." Alan F. Blackwell wrote in 1997 that "many sub-goals can be deferred to the degree that they become what is known amongst professional programmers as an 'S.E.P.'—somebody else's problem."

The belief that algorithmic bias is a dataset problem invites diffusion of responsibility. It absolves those of us that design and train algorithms from having to care about how our design choices can amplify or curb harm. However, this stance rests on the precarious assumption that bias can be fully addressed in the data pipeline. In a world where our datasets are far from perfect, overall harm is a product of both the data and our model design choices.

The goal of this article is not to convince you to ignore the data pipeline and focus solely on model design bias but rather that understanding the role that both data and the model play in contributing to bias can be a powerful tool in mitigating harm. Algorithm design is not impartial, and mitigating harm here is often more feasible than collecting comprehensive labels. Work on understanding the interactions between model desiderata and the impact of our model design choices on algorithmic bias is in its nascency. Acknowledging that model design matters has the benefit of spurring more research focus on how it matters and will inevitably surface new insights into how we can design models to minimize harm. As Lord Kelvin reflected, "If you cannot measure it, you cannot improve it."

ACKNOWLEDGMENTS

Thank you to many of my generous colleagues and peers who took time to provide valuable feedback on an earlier version of this essay. In particular, I would like to acknowledge the valuable input of Alexander D'Amour, Melissa Fabros, Aaron Courville, Hugo Larochelle, Sebastian Gehrmann, Julius Adebayo, Gregory Clark, Rosanne Liu, Hattie Zhou, and Jonas Kemp. Thanks for the institutional support and encouragement of Natacha Mainville and Alexander Popper.

REFERENCES

1. Barocas, S., Hardt, M., and Narayanan, A. (2019). Fairness and Machine Learning. https://www.fairmlbook.org.
2. Khan, Z., and Fu, Y. (2021). One Label, One Billion Faces: Usage and Consistency of Racial Categories in Computer Vision. FAccT '21:

Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 587–597, https://doi.org/10.1145/3442188.3445920.

3. Veale, M., and Binns, R. (2017). Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. Big Data & Society *4*. 2053951717743530. https://doi.org/10.1177/2053951717743530.

4. Andrus, M., Spitzer, E., Brown, J., and Xiang, A. (2021). What We Can't Measure, We Can't Understand: Challenges to Demographic Data Procurement in the Pursuit of Fairness. FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 249–260, https://doi.org/10.1145/3442188.3445888.

5. Buolamwini, J., and Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In Proceedings of Machine Learning Research, *81*, S.A. Friedler and C. Wilson, eds., pp. 1–15.

6. 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, 1711.08536 https://arxiv.org/abs/1711.08536.

7. Garg, N., Schiebinger, L., Jurafsky, D., and Zou, J. (2017). Word embeddings quantify 100 years of gender and ethnic stereotypes. Proc Natl Acad Sci U S A *115*, E3635–E3644, https://doi.org/10.1073/pnas.1720347115.

8. Hornik, R., Jacobsohn, L., Orwin, R., Piesse, A., and Kalton, G. (2008). Effects of the National Youth Anti- Drug Media Campaign on youths. Am J Public Health *98*, 2229–2236, https://doi.org/10.2105/AJPH.2007.125849.

9. Moote, A.L., and Moote, D.C. (2006). The Great Plague: The Story of London's Most Deadly Year (Johns Hopkins University Press).

10. Bagdasaryan, E., Poursaeed, O., and Shmatikov, V. (2019). Differential privacy has disparate impact on model accuracy. In Proceedings of the 33rd Conference on Neural Information Processing Systems, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, eds. (NeurIPS), p. 32. https://proceedings.neurips.cc/paper/2019/file/fc0de4e0396fff257ea362983c2dda5a-Paper.pdf.

11. Hooker, S., Moorosi, N., Clark, G., Bengio, S., and Denton, E. (2020). Characterising bias in compressed models. arXiv, 2010.03058 https://arxiv.org/abs/2010.03058.

12. Hooker, S., Courville, A., Clark, G., Dauphin, Y., and Frome, A. (2019). What Do Compressed Deep Neural Networks Forget? arXiv, 1911.05248 https://arxiv.org/abs/1911.05248.

13. Jiang, Z., Zhang, C., Talwar, K., and Mozer, M.C. (2020). Characterizing Structural Regularities of Labeled Data in Overparameterized Models. arXiv, 2002.03206 https://arxiv.org/abs/2002.03206.

14. Joseph, V., Siddiqui, S.A., Bhaskara, A., Gopalakrishnan, G., Muralidharan, S., Garland, M., Ahmed, S., and Dengel, A. (2020). Reliable model compression via label-preservation-aware loss functions. arXiv, 2012.01604 https://arxiv.org/abs/2012.01604.

About the author

Sara Hooker is a researcher at Google Brain doing deep learning research on reliable explanations of model predictions for black-box models. Her main research interests center on interpretability, model compression, and robustness.