Translation Tutorial*

Engineering for Fairness: How a Firm Conceptual Distinction between Unfairness and Bias Makes it Easier to Address Un/Fairness

Jacob Metcalf, PhD
PERVADE Team
Data & Society Research Inst.
New York, NY USA
jake.metcalf@datasociety.net

ABSTRACT

This translation tutorial will demonstrate that making fairness a tractable engineering goal requires demarcating a clear conceptual difference between *bias* and *unfairness*. Making this distinction demonstrates that bias is a property of technical judgments, whereas fairness is a property of value judgments. With that distinction in place, it is easier to articulate how engineering for fairness is an organizational or social function not reducible to a mathematical description. Using hypothetical and real-life examples I will show how product design practices would benefit from an explicit focus on value-driven decision processes. The upshot of these claims is that designing for fairness (and other ethical commitments) is more tractable if organizations build out capacity for the "soft" aspects of engineering practice.

KEYWORDS

Data ethics, algorithm ethics, machine bias, fairness, organizational ethics

ACM Reference format:

Metcalf, Jacob. 2019. Translation Tutorial: Engineering for Fairness: How a Firm Conceptual Distinction between Unfairness and Bias Makes it Easier to Address Un/Fairness In *Proceedings of Machine Learning Research. ACM FAT* Conference, New York, NY, USA, 2 pages.* https://doi.org/10.1145/1234567890

1 Description

The term "bias" connotes a very broad range of meanings. Unfortunately for the FAT* community, two very different—almost contrary—meanings collide in how we talk about ethical research and engineering in data science and algorithms. This conceptual confusion results not just in imprecise speech, but actually makes it harder to find tractable solutions to issues of fairness, transparency and accountability in data science research and applications.

The most widely used meaning of "bias," which I will call *social bias*, connotes *unfair* prejudgment of a person or group of people. In a culture with liberal democratic values, *social bias* invariably has a negative connotation. To be *socially biased* results in receiving social approbation and having your character and judgments called into question. The central judgment of *social bias* depends on the *unfairness* of prejudgment due to a principled opposition to prejudgment and an assessment of the material consequences of widespread prejudice. In other words, *unfairness* is the morally salient feature of *bias*.

On the other hand, "bias" in a statistical/technical sense connotes the gap between model and world. This form of bias is a methodological error. *Statistical bias* occurs when a model diverges from the world it is describing such that the model is ineffective for its intended purposes, usually because some population parameter was over- or under-estimated. Unlike *social bias*, *statistical bias* usually doesn't result in social approbation—it's a morally neutral determination about the effectiveness of a model.

That is unless we are discussing algorithms, where *statistical bias* becomes a reflection of and potential reinforcement mechanism for *social bias* contained in training sets. As a result, the common use of the term *bias* to connote *social unfairness* causes a conflation between *social bias* and *statistical bias* under a single term *bias*. Such terminological confusion makes it significantly harder to efficiently describe and resolve the consequences of statistical bias embedded in systems that have profound social impact. In particular, it is challenging to describe contexts where resolving *unfairness* requires introducing *statistical bias*, which is more common than anticipated.

*Article Title Footnote needs to be captured as Title Note
†Author Footnote to be captured as Author Note

WOODSTOCK'18, June, 2018, El Paso, Texas USA

In response, I argue that the FAT* community should cease to use *bias* to connote *social bias* and intentionally replace it with *unfairness*, which is already the morally salient feature of *social bias*. When these terms are disambiguated it is possible to describe *statistical bias* as a technical feature of statistical models and *fairness* as a feature of judgment about human values. Therefore, resolving *unfairness* in algorithms becomes a job for building human judgment capacity in algorithm engineering contexts.

As Arvind Narayanan [1] pointed out in his FAT* 2018 tutorial, there is no optimal or universal model of fairness to engineer for. Rather, he described 21 varieties of fairness that can be mathematically modeled, and suggested there are likely many more. Most importantly, he demonstrated that many of the headline-grabbing stories about algorithmic ethics and machine bias are at heart conflicts between types of fairness.

Given Narayanan's description of numerous varieties of fairness, *which* fairness an algorithmic system is optimized for is a matter of the organization's judgment. How this proposal adds to Narayanan's description is to point out that engineering for *which* fairness—and even deciding on *which* fairness to engineer for—requires building an organizational capacity for asking such questions. Because there is not now nor will there ever be a universally preferable model of fairness, organizations must have the capacity to explore the downstream consequences of their decision and make informed, justifiable choices.

1.1 Scenarios

I will present three scenarios to illustrate my argument.

1. **Wheelchair racing and algorithm tuning**: Imagine that you are on the computer vision team at a social networking company. Your team is tasked with designing a product that will automatically tag sporting events with descriptions and suggest stickers, filters and hashtags when a user uploads a photo or video of a sporting event, such as skiing, soccer or racing. You are working on the foot-race dataset and are currently tuning for recall (testing for false negatives). You discover that a class of racing photos tagged with typical racing-related tags, like #5k and #personalbest, are not being included in the results. What you find is that the algorithm has excluded wheelchair racing from the results, likely because the machine correlated <racing> with the use of legs.

   Such a result is *unfair* if we take seriously the value of inclusion. Nonetheless, one could argue that it is not *statistically biased* in the sense that the model does adequately represent the world, it just so happens that people who participate in wheelchair racing are not numerous enough in the learning dataset to provide a strong signal to the learning algorithms. At the very least we can assert that *statistically unbiasing* the learning set will likely not make the results more socially inclusive. Rather, the optimal solution is likely the introduction of **more** *statistical bias* through dramatically over-representing photos of wheelchair racing in the learning set to teach the algorithm to include wheelchair as a feature of the category <racing>.

2. **Hiring algorithms and gender balance**: Imagine that you are developing a product that suggests job candidates within a business networking site. The service predicts when hiring team is considering hiring for a position based on users and their colleague's activities on your service. Your product will seed suggestions for potential candidates within 3 steps on the social graph of a user's colleagues. When using an open-source machine bias tool to study your results, you realize that the algorithm is only returning 70% male candidates when the job search is for a management position. This result is possibly *statistically unbiased* because features of the learning set likely represent historical *social unfairness*—there may be fewer non-male candidates with management credentials, or it may be the case that are fewer women in the social graph of managers.

   Regardless, your team needs to decide which model of *fairness* represents your values. One model would be to simply leave the results as they are—it is arguably most *fair* to each individual job seeker to do so. On the other hand, you could hand code for *equality fairness* and force the algorithm to return results on the basis of gender demography. And yet another option is to seek *equitable fairness* and hand code to return results that dramatically favor non-male candidates in order to make up for historical unfairness. The only possible resolution to the matter involves an organizational decision about which type of fairness is desired and setting that as the value to be optimized.

3. **Recidivism scores and predictive policing**: As Narayanan argued, the infamous COMPAS debates are at heart a conflict between competing models of fairness. On the one hand, COMPAS and the justice systems that utilize it argue that fairness is best served by group parity. And in the case of COMPAS. group parity is arguable achieved—the algorithm has a similar predictive accuracy for different demographic groups. However, when critics of COMPAS, such as *ProPublica*, argue that its predictions are *biased*, what

they really mean is that the predictions are *group unfair*—the predictions cause one group to have better outcomes than another.

1.2 Organizational Capacities

In each of these cases, any technical solution is only made possible via an organizational solution. Ethical responses to machine bias will ultimately rest on the capacity of organizations that build and use algorithmic technologies to robustly articulate a model of fairness and optimize for it while moderating any negative trade-offs.

This is a commonly missed aspect of ethics in computing technology and engineering. It is not conceptually coherent to describe algorithms—nor any technology—as "ethical." Rather, the adjectival form of "ethics" should be used to describe a process, as in "ethical design practices were used in the construction of this algorithm."

Demarcating a clear conceptual boundary between *bias* and *fairness* helps make building organizational ethics capacity a more tractable task in designing for fairness, accountability and transparency. Insofar as fairness is necessarily a matter of human judgment, systematically designing for fairness increases the capacity of the organization to engage in ethical reasoning. Because there is no universal or final definition of what type of outcome is the most fair, the best criteria that we will ever have for judging whether a technology and its consequences are desirable is whether the organization that built that technology did the accounting work and has transparently shared its justifications.

Such "capacity" should be understood in a material sense that is familiar to engineering organizations already. It consists of the tools, organizational structures, and chains of accountability that are necessary to break down large engineering tasks into discrete tasks.

This tutorial will conclude with an examination of some toolkits that are currently available for building such capacity, as well as a look forward to some tools on the horizon. These toolkits will be examined in terms of thematic similarities, divergences and a consideration of what needs are currently unfulfilled.

Examples of toolkits to be considered:

- Open source/freemium
  - AuditAI (algorithmic bias tool available via GitHub)
  - Deon (ethics checklist available via GitHub)
- Corporate:
  - Facebook Fairness Flow
  - Accenture's Fairness Tool
  - Microsoft's yet-unnamed bias detection tool
- Technical Standards
  - IEEE's P7000 series
- Civil Society
  - IFF/TSSL's EthicsOS
  - Markula Center's Ethics in Technology Practice

2. Timeline

10 Min	Introduction of core argument
10 Min	Examples and analysis of conceptual confusion between fairness and bias, including video
10 Min	Discussion of case studies
10 Min	Discussion of organizational capacity and available tools
5 Min	Time cushion/Q&A
45 Min	Total

ACKNOWLEDGMENTS

Benjamin Roome assisted in the creation of some of the scenarios. Katie Shilton and Emanuel Moss provided helpful feedback during the drafting of the proposal.

REFERENCES

[1] Narayanan, A. Translation Tutorial: 21 Definitions of Fairness and Their Politics. *FAT* 2018*. https://www.youtube.com/watch?v=wqamrPkF5kk