Bias in AI: Essay

Banner ID: 000810185
Dept. of Computer Science, Durham University

I. "On Formalizing Fairness..."

How can one define fairness while evading inadvertent bias? Such is the central imperative underlying the paper "On Formalizing Fairness in Prediction with Machine Learning" [1], wherein the acute issue of safely classifying fairness with respect to machine learning is addressed. The paper illustrates the interdisciplinary nature of computer science and AI as it attempts to unify these domains with notions from the social sciences. This is achieved by surveying the various attempts at formalizing fairness within the machine learning literature and unveiling their corresponding social science counterparts.

We can not rely on computer scientists and mathematicians alone when it comes to the rapid developments within the field of AI, as the issue of *algorithmic bias* is often subtle and difficult to anticipate. More crucially, the formalization of fairness considered when endeavoring to mitigate bias must be consistent and appropriate for the associated application. Consequently, we turn our heads to the social sciences.

As asserted in [1], to devise a strictly mathematical formalization of fairness is of particular difficulty due to the interconnection with societal issues, including inequality and social conditioning, which are often difficult to quantify. Additionally, the incorporation of one specific fairness formalization within a bias mitigating machine learning model might lead to considerably different results than those attained by a near-identical model that is different only in the fairness formalization that underscores it. This precise idea is substantiated by the elaboration in [2] on the trade-offs in contrasting concepts of fairness, highlighting the incongruency between *calibration* and *equalized odds* in particular.

The main original contribution of [1] is the proposition of two new metrics for fairness within machine learning, derived from the social sciences. These are presented subsequent to extensive commentary on the multitude of other formalizations currently in use, adequately acknowledging the many limitations and drawbacks pertaining to each. An example of one well-known formalization discussed in the text is *Fairness Through Unawareness*, which is equivalent to "counter-discrimination-blind" approaches in social sciences. The approach entails the removal of protected attributes from a model's prediction process. In [3], it is contended that this measure is insufficient on account that these disused attributes may be indirectly discerned by a model from patterns existing in other attributes that are too important to discard.

Conscious of these issues together with others present in separate formalizations, [1] dispenses two prospective fairness formalizations from the social sciences that haven't yet been considered in the machine learning literature. Namely, these are *Equality of Resources* and *Equality of Capability of Function*. Whilst not providing formalizations of these, the paper is substantial in the clear and lucid exposition of the various other formalizations and its encouragement of further discussion regarding machine learning fairness formalizations that recognise social issues.

II. On the Future of Bias in AI

Beyond the popular mythos surrounding contemporary Artificial Intelligence, intelligent systems in the form of predictive models are quickly garnering an invisible omnipresence in modern quotidian life [4]. To this end, the severity of the risks facilitated by bias in such systems is uncomfortably growing.

In order to sufficiently discuss the future of algorithmic bias, we must first acquaint ourselves with its past. The problem had its first widely-reported *incident* in 1988, following the rejection of many ethnic minority applicants to a medical school purely by the decision of a newly-implemented algorithm [5]. This incident, however, is antedated by predictions made by John Weizenbaum, writing in as early as 1976 that computer programs may be afflicted with bias due to not only the data that the program uses but by the codebase itself [6], a sentiment that acts as a prescient definition of modern algorithmic bias.

While the source of the bias in the aforementioned incident was easily traceable back to the data that underscored it (that is, a dataset pertaining to past admissions, which lead to the algorithm being biased against "foreign-sounding" names), as the complexity of machine learning algorithms has grown, as has the complexity in the detection of bias within these newer models, thus the latter part of Weizenbaum's assertion is becoming increasingly pervasive. For instance, see [7], which details the eminently complex application of AI within war and weaponry. In such systems, the danger that bias posits and the sensitivity required in mitigation already vastly exceeds [5], from only 3 decades earlier.

Algorithmic bias as a sub-domain of Artificial Intelligence is something I believe is analogous with cybersecurity. Much like the perpetual need for cybersecurity, the problem of algorithmic bias is not currently, or will ever be, approaching quietus. Instead, as machine learning evolves, the ways in which bias seeps into machine learning models will evolve in parallel. Therefore, our mitigation methodologies must be continually developed, akin to the constant improvements in cybersecurity defense strategies necessitated by advances in malicious software and tactics.

To provide a concrete example, the *Generative Adversarial Network* as a general-purpose machine learning system was first introduced in 2014 [8]. Four years later, adversarial learning was introduced as a technique for the mitigation of bias [9], which was directly derived from [8]. I predict that this pattern will continue indefinitely: as new areas of bias arise, bias mitigation methods will advance by augmenting the developments within machine learning more generally.

Moving forward, within the Computer Science community, perhaps the most effective strategy that individuals can take is to spread awareness of algorithmic bias. As data and machine learning progressively integrates with many areas of the technology industry, we must not only actively ensure that bias in our products is mitigated, but also maintain a deep understanding of what it is we seek to mitigate, enabling us to be more perceptive to unexpected evolutions. Fortunately, attention given to algorithmic bias has only been growing, with resources such as [10] providing useful insights as to how those of us who often work on systems susceptible to bias can aim to achieve fairness. Of course, as [1] made clear, our knowledge of *fairness* is just as vital as our knowledge of *bias*.

REFERENCES

[1] P. Gajane, M. Pechenizkiy. (2018, May). On formalizing fairness in prediction with machine learning. *arXiv:1710.03184* [Online]. Available: http://arxiv.org/abs/1710.03184

[2] J. Kleinberg, S. Mullainathan, M. Raghavan. "Inherent trade-offs in the fair determination of risk scores," in *Proc. 8th Conf. Innovations in Theoretical Computer Science (ITCS)*, Berkeley, CA, USA, 2017, pp. 43:1-43:23.

[3] D. Pedreshi, S. Ruggieri, F. Turini. "Discrimination-aware data mining," in *Int. Conf. Knowledge Discovery and Data Mining*, Las Vegas, NV, USA, 2008, pp. 560-569.

[4] A. Pupo. (2014, July). Cognitivity everywhere: the omnipresence of intelligent machines and the possible social impacts. *World Futures Review* [Online]. 6(2), pp. 114-119. Available: https://doi.org/10.1177/1946756714533206

[5] S. Lowry, G. Macpherson, "A blot on the profession," *British Medical Journal*, vol. 296, issue 6623, pp. 657–658, Mar. 1988.

[6] J. Weizenbaum, *Computer Power and Human Reason: From Judgment to Calculation*, 1st ed. New York: W. H. Freeman and Company, 1976.

[7] J. S. Johnson, "Artificial intelligence: a threat to strategic stability," *Strategic Studies Quarterly*, vol. 14, issue 1, pp. 16-39, Feb. 2020.

[8] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio. "Generative adversarial networks," in *Conf. Neural Information Processing Systems*, Montréal, QC, Canada, 2014, pp. 2672–2680.

[9] B. H. Zhang, B. Lemoine, M. Mitchell. "Mitigating unwanted biases with adversarial learning," in *Conf. AI, Ethics, and Society*, New York, NY, USA, 2018, pp. 335-340.

[10] B. d'Alessandro, C. O'Neil, T. LaGatta. (2017, June). Conscientious classification: a data scientist's guide to discrimination-aware classification. *Big Data*. [Online]. 5(2), pp. 120-134. Available: http://dx.doi.org/10.1089/big.2016.0048