Submission for: IEEE's Technology & Society magazine

Title: Algorithmic Bias Considerations

Author: Ansgar Koene, Horizon Digital Economy Research institute, University of Nottingham, UK

In the context of the IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems, and with support from its executive director John C. Havens, Paula Boddington from the University of Oxford and myself have proposed the development of a new IEEE Standard on Algorithmic Bias Considerations. The aim is for this to become part of a set of ethical design standards, such as the IEEE P7001™ Standards Project called Transparency of Autonomous Systems with a Working Group that just started led by Alan Winfield. Whereas the Transparency of Autonomous Systems Standard will be focused on the important issue of 'breaking open the black box' for users and/or regulators, the Algorithmic Bias Standard is focused on 'surfacing' and evaluating societal implications of the outcomes of algorithmic systems, with the aim of countering non-operationally justified results.

The rapid growth of algorithm driven services has led to growing concerns among civil society, legislators, industry bodies and academics about potential unintended and undesirable biases within intelligent systems that are largely inscrutable 'black boxes' for users.

Examples which have captured the headlines include: apparent racial bias by COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) software used in various US jurisdictions to provide sentencing advice[1]; computer vision algorithms for passport photos which mistakenly register Asian eyes as closed[2]; and, beauty pageant judging algorithms that disproportionately favor white features[3]. Other examples point to gender bias such as Google's advertising algorithm which appeared to show higher paying jobs more to men than women[4].

Of importance to note, however, is that it is exceedingly rare that biased algorithm behavior is shown to have been intentional. Rather, as highlighted in the May 2016 report from the White House[5] such bias is commonly attributable to: poorly selected data; incomplete, incorrect, or outdated data; selection bias; unintentional perpetuation and promotion of historical biases [embedded in the data]; poorly designed matching systems;

[1] https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[2] https://medium.com/mit-media-lab/the-algorithmic-justice-league-3cc4131c5148#.senkns2it

[3] https://www.theguardian.com/technology/2016/sep/08/artificial-intelligence-beauty-contest-doesnt-like-black-people

[4] http://www.theverge.com/2015/7/7/8905037/google-ad-discrimination-adfisher

[5] https://obamawhitehouse.archives.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf

personalization and recommendation services that narrow instead of expand user options; decision-making systems that assume correlation necessarily implies causation; or, data sets that lack information or disproportionately represent certain populations.

Meanwhile, on the other side of the Atlantic, the EU's General Data Protection Regulation (GDPR), which was adopted in 2016 that takes effect in 2018, contains a recital (a non-binding description of the law written by its authors) stating that "a data subject has the right to an explanation of the decision reached after [algorithmic] assessment"[6]. Details of how the GDPR will be interpreted by the courts, and thus the degree to which it provides a 'right to explanation' remain to be seen.

These extensive concerns point to a clear need for ethical design standards that can help ensure that engineers, technologists, and the organizations they work for can provide clarity around how the algorithms they create deal with issues of bias in producing and in applying algorithms. Recognition of this need by the industry and research communities is clearly shown by industry initiatives such as the Partnership on AI[7], the recently launched Ethics and Governance of AI fund[8] administered by the Knight Foundation, but also investment from government sources such as the UK's Engineering and Physical Science Research Council's funding of the UnBias project[9], an initiative I co-lead in my role at the Horizon Digital Economy Research institute.

The IEEE Standards Project on Algorithmic Bias Considerations is designed to provide individuals or organizations creating algorithms with methods to provide clearly articulated accountability and clarity around how algorithms are targeting, assessing and influencing the users and stakeholders affected by the algorithm. Certification under this standard will allow algorithm creators to communicate to users, and regulatory authorities, that up-to-date best practices were used in the design, testing and evaluation of the algorithm to avoid differential impact on users that is not operationally justified. The Working Group for the Project is also open to anyone who'd like to join by getting in touch with IEEE-SA.

When properly designed, algorithmic systems provide an opportunity to help us counter existing unjustified human bias. To do so, however, needs concerted effort otherwise we run the risk that the algorithms instead entrench existing bias. This Standards Project on Algorithmic Bias Considerations is one such effort.

[6] http://fusion.net/story/321178/european-union-right-to-algorithmic-explanation/

[7] https://www.partnershiponai.org/

[8] https://www.knightfoundation.org/aifund-faq

[9] http://unbias.wp.horizon.ac.uk/