Trustworthy Algorithmic Decision-Making

[largely based on our submission to the "Algorithms in Decision Making[1]" inquiry of the UK Commons Science & Technology Select Committee]

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

UnBias[2] is a research project funded by the UK's Engineering and Physical Sciences Research Council (EPSRC grant EP/N02785X/1). The project brings together researchers from the universities of Nottingham, Oxford and Edinburgh to study the user experience of algorithm driven internet services and the process of algorithm design with special attention to the experience of young people (13 to 17 years old) and issues related to unjustified bias. UnBias aims to provide policy recommendations, ethical guidelines and a 'fairness toolkit' co-produced with young people and stakeholders from academia, teachers, NGOs, industry and regulatory organizations.

The recent research work that we have conducted with young people has highlighted important concerns around algorithm use and trust issues. Results from a series of 'Youth Juries'[3] show that many young people experience a lack of trust toward the digital world and are demanding a broader curriculum beyond the current provision of e-safety to help them understand algorithmic practices, and to increase their digital agency and confidence. Current use of algorithms in decision-making (e.g., job recruitment agencies) appears surprising to many young people. Algorithms are perceived by most of our young participants as a necessary mechanism to filter, rank or select large amounts of data but its opacity and lack of accessibility or transparency is viewed with suspicion and undermines trust in the system. The Youth Juries also facilitated young people to deliberate together about what they require to regain this trust – the request is for a comprehensive digital education as well as for choices online to be meaningful and transparent.

Bias

1. When discussing bias in algorithmic decision-making it is important to start with a clear distinction between bias that is operationally-justified and bias for which there is no justification. Justified bias prioritizes certain items/people as part of performing the desired task of the algorithm, e.g. identifying frail individuals when assigning medical prioritization. Non-operationally-justified bias by contrast is not integral to being able to do the task, and is often unintended and its presence is unknown unless explicitly looked for.
2. Given the complexities of the landscape in which algorithms are developed and used- we need to recognise that it is difficult, in some cases impossible, to develop completely unbiased algorithms and that this would be an unrealistic ideal to aim towards. Instead, it is important to base good practice on a balanced understanding and considering of multi-stakeholder needs.
3. The need for 'good practice' guidance regarding bias in algorithmic decision-making has also been recognized by professional associations such as the Institute of Electrical and Electronic Engineers (IEEE) which launched a Global Initiative for Ethical Considerations in Artificial Intelligence and

[1] https://www.parliament.uk/business/committees/committees-a-z/commons-select/science-and-technology-committee/inquiries/parliament-2015/inquiry9/
[2] http://unbias.wp.horizon.ac.uk
[3] http://oer.horizon.ac.uk/5rights-youth-juries/

Autonomous system[4] as part of which the P7003 Standard for Algorithm Bias Considerations[5] is being developed (chaired by Ansgar Koene). The validity of algorithmic system performance evaluations is by necessity limited to the specific application context for which the system was validated. In the case of algorithmic decision making that affects people, the social and cultural situation of these people forms an integral part of this context. In recognition of this context limitation, the P7003 standard is focusing on a methodological framework of identifying milestones within the algorithmic system design process that should trigger quality control procedures for finding and minimizing unjustified and inappropriate bias.

Fairness

1. The challenges in developing AI systems to deliver fair decisions is a 'wicked' problem that defies perfect solutions but demands serious attention because at its core it is not a technical problem, but rather socio-technical. It is a problem that inevitably arises from the transition of AI systems away from the structured world of 'toy challenges', like Chess and Go where the rules are undisputed, to real-world services where user expectations are imprecisely defined and the ways in which the services are used can deviate from the intentions of the developers.
2. We ran a quantitative and qualitative user study based on a resource allocation task in which participants were required to select and discuss preferred algorithms. Our survey revealed that: participants facing identical situations had different algorithmic preferences; the level of disclosed information about the algorithm significantly impacted selections; and selected preferences differed according to participant background. When discussing their preferences, participants consistently invoked normative understandings of right and wrong. Concepts of fairness were central to expressions of algorithm preference but what fairness means is contextualised and subjective, undermining the possibility of a universally preferred 'fair' algorithm. Our findings also call into question arguments that transparency can resolve fairness controversies.

Transparency

1. Meaningful transparency should also relate to a *meaningful accountability*. It is not enough for stakeholders to understand how algorithms are developed and how they make decisions, they should be given some agency to challenge algorithmic decision-making processes and outcomes.
2. In principle, algorithmic decisions can be traced, step by step, to reconstruct how the outcome was arrived at. The problem with the more complex 'big data' type processes is the high dimensionality of the underlying data. This make it very difficult to comprehend which contributing factors are salient for any given specific decision. Analytic methods for dimension reduction can often be used to make this more understandable, but may need to be applied on a case-by-case basis to appropriately evaluate important outlying and challenging cases.

[4] https://standards.ieee.org/develop/indconn/ec/autonomous_systems.html
[5] https://standards.ieee.org/develop/project/7003.html