Written evidence submitted by the Horizon Digital Economy Research Institute, University of Nottingham, and the Human Centred Computing group, University of Oxford, and (ALG0049)

1. UnBias[1] is a research project funded under the Digital Economy theme's Trust, Identity, Privacy and Security programme (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 non-operationally justified bias. UnBias aims to provide policy recommendations, ethical guidelines and a 'fairness toolkit' co-produced with young people and other stakeholders. The toolkit will include educational materials and resources to support youth understanding about online environments as well as raise awareness among online providers about the concerns and rights of young internet users. The draft report[2] summarizing the outcomes of a set of case study discussions with stakeholders from academia, teachers, NGOs and SMEs has just been finalised.

2. Professor Derek McAuley, Dr Ansgar Koene and Dr Elvira Perez Vallejos are part of Horizon Digital Economy Research[3] which is a Research Institute at The University of Nottingham and a Research Hub within the RCUK Digital Economy programme[4]. Horizon brings together researchers from a broad range of disciplines to investigate the opportunities and challenges arising from the increased use of digital technology in our everyday lives. Prof McAuley is Director of Horizon and principal investigator on the UnBias project. Dr Koene and Dr Perez Vallejos are Senior Research Fellows at Horizon and co-investigators on the UnBias[5] project. Dr Koene chairs the IEEE working group for the development of a Standards on Algorithm Bias Considerations[6].

3. Professor Marina Jirotka, Dr Menisha Patel, and Dr Helena Webb are part of the Human Centred Computing (HCC) group[7] at the Department of Computer Science, University of Oxford. This is an interdisciplinary research group that seeks to increase understanding of how innovation impacts society and advance opportunities for new technologies to be developed in ways that are more responsive to societal acceptability and desirability. Prof Jirotka, and Dr Webb are co-investigators on the UnBias project.

Questions

1. The extent of current and future use of algorithms in decision-making in Government and public bodies, businesses and others, and the corresponding risks and opportunities.

4. As part of the UnBias project we have been reviewing case studies of controversies over potential bias in algorithmic practice and scoping the informed opinion of stakeholders in this area (academics, educators, entrepreneurs, staff at platforms, NGOs, and staff at regulatory bodies etc.). It is apparent that the ever-increasing use of algorithms to support decision-making, whilst providing opportunities for efficiency in practice, carries a great deal of risk relating to unfair or discriminatory outcomes. When considering the role of algorithms in decision making we need to think not only of cases where an algorithm is the complete and final arbiter of a decision process, but also the many cases where algorithms play a key role in shaping a decision process, even when the final decision is made by humans; this may be illustrated by the now [in]famous example of the sentencing support algorithm used in some US courts which was shown to be biased[8]. Given the ubiquitous nature of computer based processing of data, almost all services, be they government, public, business or otherwise, are in some way affected by algorithmic decision-making. As the complexity of these algorithmic practices increases, so do the inherent risks of bias as there are a greater number of stages in the process where errors can occur and

accumulate. These problems are in turn exacerbated by the absence of oversight and effective regulation.

5. 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'[9] 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, especially for those unaware of such practices. Algorithms are perceived for most young people 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.

2. Whether 'good practice' in algorithmic decision-making can be identified and spread, including in terms of:

2a. The scope for algorithmic decision-making to eliminate, introduce or amplify biases or discrimination, and how any such bias can be detected and overcome?

6. When discussing bias in algorithmic decision-making it is important to start with a clear distinction between operationally-justified and non-operationally-justified bias. 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.

7. In order to identify good practice related to biases or discrimination, some important processual issues must be taken into account, for example:

   I. In order to understand the scope for algorithmic decision-making in relation to bias adequately and appropriately, it is necessary to engage with, and integrate the views of, multiple stakeholders to understand how algorithms are designed, developed and appropriated into the social world, how they have been experienced, and what the concerns surrounding their use are;

   II. Importantly, this undertaking and exploration should be achieved through rigorous research rather than abstract orientations towards good practice in relation to algorithms: thus, considering examples of the consequences that people have experienced when algorithms have been implemented, particular scenarios surrounding their use, and as emphasised in the point above- talking to people about their experiences.

   III. 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.

8. 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 in April 2016 launched a Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous system[10]. As part of this initiative Dr Koene is chairing the working group for the development of a Standard on Algorithm Bias Considerations[11] which will provide certification oriented methodologies to identify and mitigate non-operationally-justified algorithm biases through:

   I. the use of benchmarking procedures

II. criteria for selecting bias validation data sets
III. guidelines for the communication of application domain limitations (using the algorithm for purposes beyond this scope invalidates the certification)

2b. Whether and how algorithmic decision-making can be conducted in a 'transparent' or 'accountable' way, and the scope for decisions made by an algorithm to be fully understood and challenged?

9. What is essential here is to create a *meaningful transparency*: that is a transparency that all stakeholders can engage with, allowing the workings of, and practical implications of, algorithms to be accessible across the diverse stakeholder base that experience them.

10. In order to create a meaningful transparency, we need to understand what stakeholders feel such a transparency would have to incorporate for them to be adequately informed, and enable them to engage with the positive and negative implications of algorithms. Though it is unlikely that there would be complete consensus, such stakeholder engagement can provide key insights for the nature and shape of solutions to be developed.

11. Importantly, this meaningful transparency should also relate to a *meaningful accountability*. It is not enough for stakeholders just to understand how algorithms are developed and how they make decisions. In making things meaningfully transparent, stakeholders should be given some agency to challenge algorithmic decision-making processes and outcomes.

12. In principle, algorithmic decisions can be traced, step by step, to reconstruct how the outcome was arrived at. The problem with many of 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 and which are effectively acting as noise (for any given specific decision). Analytic methods for dimension reduction can be used to make this more understandable in many situations, but may need to be applied on a case-by-case basis to appropriately evaluate the important outlying and challenging cases.

13. Similarly, it is important to note that many 'big data' and 'artificial intelligence' algorithms learn from the data they are supplied with and modify their behaviour. We must look not only at the code that constitutes and algorithm, but the "training data" from which it learns. Practically this is becoming increasingly difficult as algorithms become embedded in off the shelf software packages and cloud services, where the algorithm itself is reused in various contexts and trained on different data – there is no one point at which the code and data are viewed together.

14. The IEEE Global Initiative (see point 6) are also working to establish a Standard for Transparency of Autonomous Systems[12] which aims to set out measurable, testable levels of transparency. The working group for this standard is chaired by Prof. Alan Winfield[13].

2c. The implications of increased transparency in terms of copyright and commercial sensitivity, and protection of an individual's data

15. As mentioned in our responses to 2b, while there is a need for *meaningful transparency*, this does not require that copyrighted code (or data) is made public. Within the community currently researching this topic, a recurring suggestion is the use of a neutral (or government associated) auditing body that could be tasked with certifying algorithmic systems through a process of expert analysis. This algorithm auditing could be done under a non-disclosure-agreement, protecting the IP, and the individual data. A detailed discussion outlining arguments in favour of such an approach was developed in an open access published paper by Andrew Tutt with the title "An FDA for Algorithms"[14].

16. Even if the copyrighted code is not made public, somehow making aspects of the design of algorithms more visible may still be useful. We see how the food industry make elements of their produce accessible for consumers to allow for consumers to make informed decisions about what they purchase. At this point it is difficult to say what is better/worse without full and proper engagement with industry and other stakeholders, as we are currently engaged in through the UnBias project.

17. It is necessary to have a dialogue with industry to understand their genuine concerns surrounding increased transparency, and how a way forward can be forged. There are elements of business procedures which have to be made transparent already (e.g. the requirements for audit, health and safety, etc…) so it is not that they are unaccustomed to such requirements. However, given that there is an element of commercial sensitivity in this context, then it is important to see what suggestions they would have to allow for increased transparency.

18. We should be careful that we do not give the impression that commercial interests supersede the rights of people to obtain information about themselves. We should be cautious about assuming industry interests are more important than other ones, and move forward with a balanced approach.

19. Finally, the traditional bargain between society and inventors has been the patent - disclosure to stimulate innovation in return for commercial protection – the question arises as to what role might patents play in transparency. However, the situation concerning software patents is globally complex, but then the issue of algorithmic transparency is rapidly becoming a global issue.

3. Methods for providing regulatory oversight of algorithmic decision-making, such as the rights described in the EU General Data Protection Regulation 2016

20. The right to explanation in GDPR is still open to interpretation and the actual practice will become established as cases unfold when enforcement starts in 2018. For example, the right to recourse and to challenge algorithmic made decisions, is restricted to decisions that are made *fully autonomously* by algorithms and that have clearly *significant* impact on the person – it will be some time before we understand how these clauses will be implemented, and with impending Brexit, whether the UK will continue to align with the EU on these interpretations. The recent paper by Wachter et al.[15] puts forward the case that much more is needed to deliver a '*right to explanation'*.

21. More broadly, it is our position as a project that open dialogue amongst key stakeholders is an important step towards advancing the responsible oversight of algorithmic decision-making. It is necessary to include the perspectives of those from a wide range of sectors, alongside government and industry, in order to scope concerns over the current and future use of algorithms, and to identify genuine opportunities for regulation that are both technically feasible and legally and societally valid. As noted above, the activities of the UnBias project include the scoping of opinion amongst a wide range of informed stakeholders. By promoting discussion between stakeholder groups we are working to identify potentially effective methods for oversight of algorithmic decision- making. From the work we have conducted in this area so far, it is clear (as described above) that transparency alone is not a meaningful solution to the potential problems caused by algorithmic practices. Regulatory oversight needs also to incorporate responsibility and accountability so that users affected by algorithmic-decision making have opportunities to 1) understand how decisions about them were reached and 2) challenge those decisions if they feel them to be unfair. As also noted above, suggestions emerging from our project stakeholder dialogue so far include the possibility of an expert auditing or ombudsman system that oversees practice and mediates disputes. Further suggestions, in line with developments by the IEEE and elsewhere, include the provision of industry standards and certificates.

22. The Council of Europe's Committee of Experts on Internet Intermediaries (MSI-NET)[16] is currently also exploring the human rights dimensions of automated data procession techniques (in particular algorithms) and possible regulatory implications. As part of this investigation a preliminary report[17] was published on February 20th which includes a number of relevant case studies and recommendations that are applicable to the topic of this inquiry.

April 2017

---

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

[2] http://unbias.wp.horizon.ac.uk/wp-content/uploads/2016/08/UnBias_Stakeholder_1stWorkshop_report_draft_for_approval.pdf

[3] http://www.horizon.ac.uk

[4] https://www.epsrc.ac.uk/links/councils/research-councils-uk-rcuk/digital-economy-research-rcuk/

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

[6] https://standards.ieee.org/develop/project/7003.html

[7] https://www.cs.ox.ac.uk/activities/HCC/

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

[9] http://oer.horizon.ac.uk/5rights-youth-juries/

[10] https://standards.ieee.org/develop/indconn/ec/autonomous_systems.html

[11] https://standards.ieee.org/develop/project/7003.html

[12] https://standards.ieee.org/develop/project/7001.html

[13] http://people.uwe.ac.uk/Pages/person.aspx?accountname=campus\a-winfield

[14] https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2747994

[15] https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2903469

[16] https://www.coe.int/en/web/freedom-expression/committee-of-experts-on-internet-intermediaries-msi-net-

[17] http://rm.coe.int/doc/09000016806fe644