Executive Summary

- This submission outlines when and how discrimination may be introduced or amplified in algorithmic decision-making.
- The various stages in the algorithmic decision-making process offer opportunities for the regulation of algorithms to prevent and/or address potential discrimination.
- Understanding the complexities involved in algorithmic decision-making and challenges for regulation is essential for the development of an appropriate and effective regulatory framework.

Introduction

1. This submission is made by the Human Rights, Big Data and Technology Project (‘the Project’).¹

2. The Human Rights, Big Data and Technology Project, funded by the Economic and Social Research Council and based at the University of Essex’s Human Rights Centre, analyses the challenges and opportunities presented by the use of big data and associated technologies from a human rights perspective. Drawing on the wide range of expertise of its interdisciplinary researchers and partner organisations, the Project considers whether fundamental human rights concepts and approaches need to be adapted to meet the rapidly evolving technological landscape. The work also brings together practitioners, States, industry, and United Nations’ officials, and academics in the fields of human rights, big data and associated technologies to assess existing regulatory responses and whether reforms are needed in order to maximise effective human rights protection.

3. The Project is grateful for the opportunity to present this submission to the Science and Technology Committee. The submission is drawn from the Project’s work on the human rights implications of algorithmic decision-making. This research focuses on the development of approaches to prevent, identify and remedy discrimination in algorithmic decision-making, and questions how algorithms may be developed and used so as not to threaten or violate human rights. Within this submission, we focus on the conditions under which discrimination may be introduced or amplified within the algorithmic decision-making process and outline possible ways in which the design, testing and assessment of algorithms can be regulated in order to address the potential for discrimination.

The Human Rights Implications of Algorithmic Decision-Making

4. Algorithmic decision-making is not new.² Algorithmic decision-making processes can be semi- or fully-automated, which varies the extent of human influence in the actions that follow from the outcomes an algorithm generates.³ Algorithms are deployed across public and private life

---

¹ The Human Rights, Big Data and Technology Project, available at <http://www.hrbdt.ac.uk>
² As a method or technique for making decisions, algorithms are not new. Algorithms have been used for a long time to solve problems and complete tasks, which range from the calculation of (predicted) polling results, to email systems and online search engines that are widely used today. Algorithmic decision-making examples that predate computer-based testing include psychometric testing for interview selection and intellectual quotient tests for school admissions.
³ This distinction has surfaced in recent legislation such as the Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)
Written evidence submitted by The Human Rights, Big Data and Technology Project (ALG0063)

and often offer significant benefits in relation to efficiency and computational power, that can facilitate decision-making.⁴

5. At the same time, discrimination can be introduced and amplified by algorithmic decision-making as documented in a number of recent reports.⁵ A series of reports commissioned by former US President Obama studied the effects of big data and algorithms,⁶ one of which highlighted the importance of examining how algorithmically-driven decisions might exacerbate existing socio-economic disparities and concluded that the use of algorithms should be monitored for potential discriminatory outcomes.⁷ Similar sentiment has been expressed in a recent report by the UK Information Commissioner’s Office.⁸ This report raised concerns about the discriminatory effects of algorithmic decisions and recommended a preventative approach, emphasizing the need to detect discrimination in algorithmic systems.⁹

6. The potential dangers of algorithmic discrimination are also often raised in relation to criminal sentencing. For example, investigative journalists from ProPublica studied the risk scores for people arrested in Broward County in Florida, generated by algorithms used in criminal sentencing. They compared the predicted recidivism rates to actual recidivism,¹⁰ concluding that there was a high rate of error in the results generated by the algorithm and a significant racial disparity biased against black defendants. Such algorithmic assessment tools are used widely, in other states in the US and other countries.

7. In particular, the software ‘COMPAS’ that ProPublica investigated is also used in Wisconsin, where it has been challenged in legal proceedings.¹¹ A defendant appealed against the reliance on the algorithmic assessment in his sentencing decision. In its decision, the Wisconsin Supreme Court focused on whether the tool was used properly – whether limitations

('GDPR') [2016] OJ L119/2, art. 2(1).


⁹ Big data, artificial intelligence, machine learning and data protection, p52-53.

¹⁰ Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ‘Machine Bias’ (ProPublica, 23 May 2016), available at <www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>

and cautions were observed. In doing so, it rejected the defendant’s appeal, although the conclusion also implied that COMPAS should not be determinative in sentencing decisions.\textsuperscript{12}

8. These reports should not necessarily lead to the conclusion that algorithmic decision-making is inherently problematic or imply that human decision-making is not susceptible to discrimination, bias or unfairness. What they should underscore, however, is that algorithmic decision-making is not neutral as is often assumed. While, in theory, mathematics itself is neutral, mathematical algorithms are formulated based on a set of assumptions and data inputs (amongst other factors), which may not be objective and can embed bias.\textsuperscript{13} Consequently, just as humans can discriminate, so too can algorithms.\textsuperscript{14} Algorithmic decision-making not only carries the risk of introducing discrimination but also of amplifying existing discriminatory approaches under the guise of neutrality. The appearance of neutrality within algorithmic decision-making may make the detection of discriminatory elements difficult since algorithmic systems can create less visible forms of discrimination.\textsuperscript{15}

9. Recognition of the discriminatory potential of algorithmic decision-making requires a regulatory response that focuses on preventative approaches to discrimination,\textsuperscript{16} particularly in the design phase, as well as internal and external oversight mechanisms, human rights impact assessments\textsuperscript{17} and processes designed to identify and eliminate discrimination and provide remedies to individuals and groups affected.

10. A regulatory approach should also take into account the diverse range of algorithms utilised, and the increasing sophistication and complexity of algorithms that are now generated by (semi)-automated processes rather than hand-written. Attempting to future-proof regulatory responses will therefore be critical to their effectiveness in addressing the potential for discrimination in algorithmic decision-making.


\textsuperscript{14} It should be noted here that the concern about the potential for discrimination here refers to the understanding and prohibition of discrimination in human rights, referring to “any distinction, exclusion, restriction or preference which is based on any ground such as race, colour, sex, language, religion, political or other opinion, national or social origin, property, birth or other status, and which has the purpose or effect of nullifying or impairing the recognition, enjoyment or exercise by all persons, on an equal footing, of all rights and freedoms”, as per the United Nations Human Rights Committee’s interpretation on non-discrimination in the International Covenant on Civil and Political Rights, in its General Comment 18, para 7. This differs from the inherent nature and purpose of algorithms to discriminate, in the computational sense of distinguishing between items in classification models.


The Potential for Discrimination to be Introduced or Amplified within the Decision-Making Process

11. The development of a regulatory approach to prevent and mitigate the risks of discriminatory effects requires an understanding of how discrimination can arise within the algorithmic decision-making process. This section focuses on when and how discrimination can occur as a baseline for preventing and identifying discrimination.

12. Discrimination can be introduced and built into algorithms at discrete stages, and any existing discrimination may also be amplified as it flows through the algorithmic decision-making process. To identify and address the points at which discrimination may be introduced or amplified, the process must be broken down. Specific questions include:

   (1) What are the stages involved in the algorithmic decision-making process;
   (2) What are the potential discriminatory risks at each stage;
   (3) What are the interconnections between these problems;
   (4) To what extent can they be isolated?

Stages in the algorithmic decision-making process generally include the following:

13. The algorithmic decision-making process as presented in the figure above, may be entirely internal (e.g. within a corporate structure), or alternatively, some stages may be conducted internally and others contracted to third party actors. For example, an organisation may conceptualise an issue and conduct initial analysis to determine the purpose of the algorithm, but the design and testing phases may involve the purchase of datasets and data science solutions, and monitoring and evaluation of the performance of the algorithm may be assessed by an independent third party.

**Conceptualisation and initial analysis**

14. An algorithm is typically designed to solve a specific problem, and algorithms are used extensively across a large spectrum of activities to solve various problems in everyday life. This stage involves the identification of an issue or a problem to be addressed, and initial analysis to determine how resolution of the issue or problem can be achieved. In this process, the purpose of the algorithm – the specific problem it is intended to solve – and the context for its deployment is thought through. For example, smoke alarms are used to detect fire (the problem), but do so by examining contextual factors, in this case the level of smoke present in a room.
Design and Testing

15. The design and testing phases are iterative processes, refined through repeated testing and design modification. If discrimination is introduced at the design and testing phase, subsequent deployment of the algorithm can amplify discriminatory effects.

16. To be effective, algorithms rely on accurate and good quality data inputs.\(^{18}\) The variables used in the design of the algorithm determine the factors that are taken into consideration and those that are excluded, as well as their relative weight in the calculation of the outcome. If the wrong inputs are selected, or if the parameters are inaccurate, the outcome may be flawed and potentially meaningless. To put it simply, predicted outcomes might not be meaningful if the data inputs are inaccurate, unreliable, or unrepresentative. Inputs should be tested to ensure that they are appropriate to measure the desired outcome, and accurately reflect the context in which the algorithm will be deployed.

17. For example, crime mapping tools are increasingly used to examine crime data and to identify hotspots to increase efficiency in the allocation of police resources. However, the input data (e.g. historical crime data) can be skewed by policing practices that may not reflect the actual incidence of crime, and are instead influenced by the targeting of marginalized groups and over-policing of particular areas. If the data used are not scrutinized for quality and are in fact not reflective of the context in which an algorithm may be deployed, discrimination can be introduced through biased data, skewing the outcomes.

18. A potential problem may arise at this stage if the best possible data input that offers the true measure (or close to it) is unavailable, inaccessible to the designers of the algorithm, or impossible to quantify for the purposes of the algorithm. In such situations, proxies – alternative inputs to represent other unavailable or inaccessible data in a calculation – are often used as substitute variables. However, the use of proxies poses certain problems as they can misrepresent reality.

19. For example, to increase the efficiency of hiring decisions, some companies have turned to solutions such as automated filtering of job applications.\(^{19}\) These algorithmic decision-making processes compare potential candidates against the ideal profile for the position. In the calculation of relevant factors, on the assumption that it is more likely that someone with a long commute will look for alternative, more convenient, employment, commute distance has been used to proxy for the risk of someone quitting their job: the further they live from the location of work, the more likely they will tire of the commute and quit. This is, however, problematic as residential location can also proxy for a candidate’s income levels – those who live further away from jobs in the city centre do so partly due to the lower cost of rent and cost of living in general. In such case, discrimination against already marginalised communities is introduced in hiring decisions and can contribute towards perpetuating inequality.\(^{20}\)

---

\(^{18}\) Jeff Larson, Julia Angwin and Terry Parris Jr., ‘How Machines Learn to be Racist’ (ProPublica, 19 October 2016), available at <https://www.propublica.org/article/breaking-the-black-box-how-machines-learn-to-be-racist>


20. The challenge in ensuring that input data is of sufficient quality is increased when the design of the algorithm is outsourced to external data science solution providers, or if the datasets are procured from third parties such as data brokers. Data brokers – companies that collect personal information about consumers and resell or share that information – are a key supplier in the information economy, providing data inputs for algorithmic decision-making processes. Concerns about the transparency and accountability of such companies have been raised in this regard, as these business-to-business enterprises often do not have direct consumer interaction.  

*Deployment*

21. A key question at this stage is whether the context in which the algorithm is deployed and the purpose for its use, matches the purpose and the context for which it was designed. Beyond the design and testing of the robustness of a ‘prototype’, discriminatory effects can arise even when an algorithm is deployed in the intended context.

22. A significant draw in the deployment of algorithms today is the availability of big data, that increases the scale, computational power and efficiency of data analytics. This raises a key issue of foreseeability, if previously invisible or unpredictable connections are made available through analytics. Beyond the challenge in mapping foreseeable impact and addressing any discriminatory outcomes, the greater challenge is how such unforeseeable impact should be treated and managed. There is a lack of tenable solutions to this problem.

*Recommendations: Transparency, Accountability and Regulation*

23. The foregoing demonstrates the discriminatory potential associated with algorithmic decision-making and the need to develop approaches to prevent and identify discrimination at each stage in the algorithmic decision-making process. This section outlines some parameters for consideration at each decision-making stage. These may offer opportunities to mitigate the risk of discrimination.

24. Further work in relation to how semi- or fully-automated algorithmic decision-making processes can be effectively regulated requires consideration of the possibilities and limits of regulation, including the feasibility of self-regulation by machines. At present, limitations on the automated processing of personal data, such as in the GDPR, prescribe special protections around ‘sensitive data’. However, information that might not be obviously sensitive, such as metadata, can be used to derive further conclusions about individuals when combined with other data in algorithmic analysis. As such, more work is needed to improve meaningful protection of individuals against rights abuses arising from algorithmic decision-making. Attention must also be paid to the role played by humans in the algorithmic development and monitoring process. Greater clarity is needed on the interaction between manual human evaluation and automated algorithmic systems, and how safeguards can be implemented.

---


22 ‘Metadata’ (Privacy International), available at <https://www.privacyinternational.org/node/53>


*Error! Hyperlink reference not valid.*
25. In this submission, the HRBDT project suggests that questions of transparency and accountability need to be considered at each stage in the algorithmic decision-making process and in the process as a whole in order to offer effective solutions. Transparency in the testing methodology of the algorithmic decision-making process is particularly important for systems that are used to implement public policy. As many scholars, policy-makers and practitioners have already argued, key principles such as transparency and accountability for data processing practices are critical measures. At the same time, debates persist on how to operationalize principles such as transparency against increasingly complex algorithms and opaque decisions, and to ensure accountability for adverse consequences such as discrimination. Debates also continue into how transparency can be ensured without requiring business enterprises to reveal their full algorithms, thereby losing their proprietary privilege. Notwithstanding these debates into practical implementation issues, principles of transparency and accountability are a crucial part of identifying problems and achieving accountability. Sophisticated responses to these challenges will be required in order to address the risks to human rights posed by algorithms.

26. To address discrimination, there is an implicit question about what algorithmic fairness means, how one derives the basis of fairness and how fairness can be measured. It is proposed that the design and testing of algorithms should be approached with a view to prioritising the prevention or at least minimization of discrimination in outcomes. This could yield different results than using abstract proxy metrics for distinct groups to assess discrimination theoretically.

27. At the stage of conceptualisation and initial analysis, thought needs to be given to the purpose of the algorithmic system, and the broader issue or problem it is intended to address. If the purpose is itself discriminatory, this should immediately preclude the algorithm from being developed and/or deployed. Another situation may be that classification algorithms that are not obviously unlawful but designed to differentiate and sort data in a certain way could reveal a propensity for indirect discrimination when tested. If the discriminatory effects, including


27 Classification algorithms are one category of algorithms. Different types of algorithms can be designed to fulfill various functions, such as to draw connections between information to find associations and patterns, filter through information to extract certain results, and sort and rank results by prioritisation. See also, Nicholas Diakopoulos, ‘Algorithmic Accountability: Journalistic investigation of computational power structures’ (2015) 3(3) Digital Journalism
indirect discrimination, are not resolved, the algorithm should not be deployed. It is also critical at this phase to conduct initial analysis on the relationship between the purpose of the algorithm and the context for which it is intended to be deployed, and assess the likely impacts of deployment on this context. If likely impacts include some direct or indirect discriminatory effect, solutions to eliminate the discriminatory impact must be explored in order to proceed. For example, the design of the algorithm could include conditions to exclude such situations.

28. At the design and testing phase, data inputs must be examined, refined, updated and re-examined to ensure the quality of the data used and the accuracy of the algorithm. In the case of data inputs being procured from third parties such as data brokers, even if access to the original basis on which the data was collected is unavailable, a process should be developed whereby the quality of input data is evaluated.

29. To ensure that the algorithm is reliable, the outputs generated should be monitored and tested against control groups for accuracy in expected results. Even if an algorithm does not flag issues in individual test cases, repeated testing can reveal trends in the outcomes, and certain parameters that correlate with particular results. In this case, even if discrimination were not embedded in the data used, indirect discrimination could be a problem, unless detected in the testing phase and the feedback returned to the design phase for modification of the algorithm.

30. Prior to deployment, testing for any foreseeable impact of the algorithm can mitigate such impact if counter-measures are devised to respond to or prevent any discriminatory impacts foreseen. The evaluation at this stage should also consider if foreseeable impacts may preclude deployment.

31. Consistent and repeated monitoring and evaluation over time is important to ensure that the performance of the algorithm is as expected. This is critical as sample bias might not be initially obvious and results may be only provisionally accurate. Changes such as those relating to environmental factors can also affect the strength of the assumptions for which the algorithms were purposed and designed. If anomalies are not detected and interrogated, this can result in adverse consequences including indirect discrimination. This monitoring and evaluation over time should also monitor for indirect discrimination, and any such assessment by an independent third party should be facilitated with access to relevant data inputs, design and testing information for an effective evaluation.

32. Human rights impact assessments are critical to test and assess the human rights impact at each stage of the algorithmic decision-making process, which should at a minimum be conducted internally. Additionally, it is preferable if human rights impact assessments are also conducted by external auditors, such as human rights experts. One option that has been introduced in the discourse about regulatory oversight is the possibility of black-box testing, where external assessors have the opportunity to analyse the overall system behaviour of algorithms, if some details of the algorithms are not made available. In the case of external oversight mechanisms, access to all relevant information, assumptions, processes and stakeholders to assessors is critical to conducting a meaningful human rights impact assessment. In general, a strong preventative approach is recommended. The basis for testing the human rights impact of the algorithmic decision-making should be grounded in the
international human rights legal framework, with clarity in the scope of rights included in the assessment and the way in which standards are applied.

33. Acknowledging that there may be limitations in preventing discrimination and other adverse impacts arising from algorithmic decision-making processes, effective remedies are critical to provide recourse to individuals or groups whose rights have been affected. This is aside from, and different to, rectifying technical problems in algorithms.

April 2017