Centre for Data Ethics and Innovation

Centre for Data Ethics and Innovation: Review on Bias in Algorithmic Decision Making

Terms of Reference: Bias in Algorithmic Decision Making

Machine learning algorithms often work by identifying patterns in data and making recommendations accordingly. This can be supportive of good decision making, reduce human error and combat existing systemic biases. However, issues can arise if, instead, algorithms begin to reinforce problematic biases, either because of errors in design or because of biases in the underlying data sets. When these algorithms are then used to support important decisions about people's lives, for example determining their credit rating or their employment status, they have the potential to cause serious harm.

The CDEI will investigate if this is an issue in certain key sectors, as well as the extent of this, and produce recommendations to Government, industry and civil society about how any potential harms can be minimised.

What is the issue?

The UK has a robust legislative framework to ensure people are treated fairly and are not subject to discrimination. Despite these legal protections, there is evidence that biased decision-making, both conscious and unconscious, continues to exist across many sectors. This can bring with it substantial missed opportunities both for individuals, who may experience unfair treatment, as well as organisations, where biased decision-making processes can result in worse outcomes. Algorithmic decision making offers an opportunity to challenge these biases. In designing these systems we can actively seek to identify and eradicate unfair and arbitrary decision making and address the biases which humans bring with them in assessing scenarios. However, if machine learning algorithms are used to support human decision making, or even make decisions independently, we must be confident that they themselves are making those decisions in a way which we consider to be justifiable.

Some patterns identified by algorithms may act as legitimate guides to decision making, for example, a credit algorithm might refuse a loan to someone with a history of defaulting on debts. However, when algorithms begin to make decisions based on attributes which may act as proxies for other demographic characteristics, in particular when making decisions which reinforce pre-existing social inequalities (for example, an algorithm which exhibits racial bias in making parole decisions), they become harmful.

What will our focus be?

We will take a sectoral approach to this Review because, in our view, the ethical dimensions of the decisions which algorithms make cannot be disentangled from the context in which they are being made. Considerations which might be fundamental to an algorithm setting

insurance premiums, could be absolutely irrelevant when using similar technology to predict an individual child's risk of abuse or neglect.

We have chosen to select four sectors which involve both decision making with a high potential impact on individuals, and historic evidence of bias. These sectors are:

- **Financial services:** Our focus is on the use of machine learning and AI in credit and insurance decisions about individuals and the ability of financial services companies to identify and mitigate potential biases as they invest in these new technologies.
- **Crime and justice:** The use of predictive algorithms in both policing and judicial decision making is a growing field, in particular in the US, but now also increasingly in the UK. Existing bias in policing and justice is well documented, and inaccurate and unfair decision making in this sector can have enormous consequences for individuals and society.
- **Recruitment:** Machine learning may offer significant potential to speed up and improve the quality of recruitment processes. However, historic recruiting patterns were often significantly biased by today's standards and it is key that new software is designed and implemented in a way which challenges rather than embeds these inequalities.
- **Local government:** There is early research to suggest that algorithmic decision making may have a role to play in allocation of local government resource, in particular in identifying cases of potential child abuse and neglect which should be escalated. Clearly, this is an area of the highest sensitivity and, if these technologies are to be adopted, it must be done with the highest ethical standards.

How will we work?

We plan to engage with stakeholders across the chosen sectors to build an understanding of current practice. We will remain flexible on our methodology and vary our approach between sectors, depending on what approach best fits the individual contexts. We aim to support the development of technical means for identifying algorithmic bias that have scope to be applied across the chosen sectors, and produce recommendations to government and the relevant sectors about how any potential harms can be identified and minimised.

We will phase our work on the different sectors, focussing on financial services and crime and justice from April 2019-September 2019 and then on recruitment and local government from October 2019-March 2019.

What types of outputs will we produce?

Examples of the types of outputs we may produce for specific sectors include: operational codes of practice for the trialling of algorithmic decision making tools, bias tests which can be used by companies to identify and mitigate bias in their own algorithms, and procurement guidelines to be followed by those buying algorithms from technology providers.

Centre for Data Ethics and Innovation

Each sector will be addressed with a view to the specific needs of the sector involved, as well as the broader applicability of the findings and outputs. Our final report will help us do this by synthesising these different elements, summarising work done in each sector, and drawing out tools and findings which are applicable to tackling bias in algorithmic decision making more generally.

What are our timelines?

An interim report will be published by Summer 2019, and a final report, including recommendations to government, by March 2020.

Open call for evidence

As part of the Review into Bias in Algorithmic Decision Making, we are taking submissions via an open call for evidence in relation to four key sectors: financial services, crime and justice, local government and recruitment.

We are particularly interested in hearing from a broad range of stakeholders working in or specialising in any of the four sectors. These include: frontline public servants (police officers, social workers and council officers); banks, recruitment, insurance, and financial services companies; academics, data scientists, research and policy organisations; technology companies, start-ups, providers, developers, and buyers; regulators, inspectors, professional standards and ethics bodies; groups representing minority and/or disadvantaged groups; privacy activists, civil liberty organisations and rights campaigners.

We encourage members of the public, who may not clearly fit into these categories, but who may have been personally impacted by an algorithmic decision making process, to get in touch. As part of the wider Review we will also identify opportunities to engage meaningfully with the public, in particular those who could be at higher risk of being disadvantaged by the growing use of algorithms, on these issues to ensure their views feed into our policy recommendations.

Please respond to the questions below in reference to one (or more) of our four key sectors: **financial services, crime and justice, local government, and recruitment.**

1.The use of algorithmic tools:
1.1 What algorithmic tools are currently being developed or in use?
1.2 Who is developing these tools?
1.3 Who is selling these tools?
1.4 How are these tools currently being used? How might they be used in the future?
1.5 What does best practice look like and are there examples of especially innovative approaches?

Centre for Data Ethics and Innovation

1.6 What are the key ethical concerns with the increased use of algorithms in making decisions about people?

2. Bias identification and mitigation:
2.1 To what extent (either currently or in the future) do we know whether algorithmic decision making is subject to bias?
2.2 At what point is the process at highest risk of introducing bias? For example, in the data used to train the algorithm, the design of the algorithm, or the way a human responds to the algorithm's output.
2.3 Assuming this bias is occurring or at risk of occurring in the future, what is being done to mitigate it? And who should be leading efforts to do this?
2.4 What tools do organisations need to help them identify and mitigate bias in their algorithms? Do organisations have access to these tools now?
2.5 What examples are there of best practice in identifying and mitigating bias in algorithmic decision making?
2.6 What examples are there of algorithms being used to challenge biases within existing systems?

3. Public engagement:
3.1 What are the best ways to engage with the public and gain their buy in before deploying the use of algorithms in decision making? For example, should a loan applicant be told that an algorithm is being used to assess their loan application?
3.2 What are the challenges with engaging with the public on these issues?
3.3 What are good examples of meaningful public engagement on these issues?

4. Regulation and governance:
4.1 What are the gaps in regulation of the use of algorithms?
4.2 Are there particular improvements needed to existing regulatory arrangements to address the risk of unfair discrimination as a result of decisions being made by algorithms?

We welcome written submissions in Word format of 2,000 words or under to policy@cdei.gov.uk.

The deadline for responses on crime and justice and financial services is **14 June 2019**.

The deadline for responses on local government and recruitment is **19 July 2019**.

While we are particularly keen to receive evidence specifically relating to the four key sectors outlined above, if you have evidence regarding bias in algorithmic decision-making more generally, then we would also be happy to receive these submissions. The deadline is **14 June 2019.**

We will publish a summary of responses over the Autumn.

In your response, please clarify:

- If you are responding on behalf of an organisation or in a personal capacity
- Whether you are responding to the call for evidence for the algorithmic bias review, the online targeting review, or both
- Which questions you are answering by referring to our numbering system. There is no need to respond to all of the questions if they are not all relevant to you
- Whether you are willing to be contacted (in which case, please provide contact details)
- Whether you want your response to remain confidential for commercial or other reasons. If you prefer to engage in person please specify this. We will try our best, resource-allowing, to find opportunities to do this.

Information provided in response to this call for evidence, including personal information, may be published or disclosed in accordance with the access to information regimes (these are primarily the Freedom of Information Act 2000 (FOIA), the General Data Protection Regulations (GDPR), and the Environmental Information Regulations 2004).

If you want the information that you provide to be treated as confidential, please be aware that, under the FOIA, there is a statutory Code of Practice with which public authorities must comply and which deals, amongst other things, with obligations of confidence. In view of this it would be helpful if you could explain to us why you regard the information you have provided as confidential.

If we receive a request for disclosure of the information we will take full account of your explanation, but we cannot give an assurance that confidentiality can be maintained in all circumstances. An automatic confidentiality disclaimer generated by your IT system will not, of itself, be regarded as binding.

We will process your personal data in accordance with the GDPR.