July
2019

Interim report

Review into bias in algorithmic decision-making

Centre for
Data Ethics
and Innovation

About the CDEI

The adoption of data-driven technology affects every aspect of our society and its use is creating opportunities as well as new ethical challenges.

The Centre for Data Ethics and Innovation (CDEI) is an independent advisory body, led by a board of experts, set up and tasked by the UK Government to investigate and advise on how we maximise the benefits of these technologies.

The CDEI has a unique mandate to make recommendations to the Government on these issues, drawing on expertise and perspectives from across society, as well as to provide advice for regulators and industry, that supports responsible innovation and helps build a strong, trustworthy system of governance. The Government is required to consider and respond publicly to these recommendations.

We convene and build on the UK's vast expertise in governing complex technology, innovation-friendly regulation and our global strength in research and academia. We aim to give the public a voice in how new technologies are governed, promoting the trust that's crucial for the UK to enjoy the full benefits of data-driven technology.

The CDEI analyses and anticipates the opportunities and risks posed by data-driven technology and puts forward practical and evidence-based advice to address them. We do this by taking a broad view of the landscape while also completing policy reviews of particular topics.

In the October 2018 Budget,[1] it was announced that the CDEI would be exploring the use of data in shaping people's online experiences and the potential for bias in decisions made using algorithms. These two large-scale Reviews form a key part of the CDEI's 2019/2020 Work Programme.[2]

More information about the CDEI can be found at www.gov.uk/cdei

[1] www.gov.uk/government/publications/budget-2018-documents/budget-2018

[2] www.gov.uk/government/publications/the-centre-for-data-ethics-and-innovation-cdei-2019-20-work-programme/the-centre-for-data-ethics-and-innovation-cdei-2019-20-work-programme

Contents

Foreword

Roger Taylor
Chair, Centre for Data Ethics and Innovation

New data-driven technology is transforming our society. Whether it is deciding what video to recommend or diagnosing a serious illness, the ability of machines to process vast amounts of data and make decisions is powering the economy and industry, reshaping public services and opening up new areas of research and discovery.

Artificial intelligence and algorithmic systems can now operate vehicles, decide on loan applications and screen candidates for jobs. The technology has the potential to improve lives and benefit society but it also brings ethical challenges which need to be carefully navigated if we are to make full use of it. It is an issue that governments worldwide are now grappling with – including the UK through the establishment of the CDEI. There are significant rewards for societies that can find the right combination of market-driven innovation and regulation to maximise the benefits of data-driven technology and minimise the harms. The UK, with its robust legal and regulatory systems, its thriving technology industry, and its leading academic institutions is well placed to achieve this.

Our goal is an environment in which the public are confident their values are reflected in the way data-driven technology is developed and deployed, where we can trust that decisions informed by algorithms are fair, and one where risks posed by innovation are identified and addressed. It is in such an environment that ethical innovation will flourish.

The CDEI's role is to set out what needs to be done to address the challenges and realise the benefits posed by data-driven technology. We will do this by providing high quality and robust advice to the Government. Our role is in part to identify gaps in current governance frameworks, whether that is in legislation, regulation or other mechanisms. But equally important is our duty to identify how public policy can support the development of technologies and industries which allow us to benefit safely from automated decision systems, robotics and artificial intelligence.

While data-driven technology continues to develop at great speed there is no shortage of predictions about its future impact. But some of the challenges are with us now and are not merely theoretical. In these Reviews, the first for the CDEI, we are focusing on two of the more urgent issues – Online Targeting and Bias in Algorithmic Decision-Making. Both are topics which cut across a range of applications of data-driven technology and force us to confront different ethical questions.

I am delighted to publish our interim reports setting out the progress we have made in our first two Reviews. The reports outline our analysis to date and our emerging insights as we develop our recommendations for the Government.

I would like to thank all those who have inputted to the CDEI's work so far. These contributions have been invaluable in helping us explore the complicated issues we need to understand. I look forward to continuing to work together as we develop our final recommendations.

Executive summary

The use of algorithms has the potential to improve the quality of decision-making by increasing the speed and accuracy with which decisions are made. If designed well, they can reduce human bias in decision-making processes. However, as the volume and variety of data used to inform decisions increases, and the algorithms used to interpret the data become more complex, concerns are growing that without proper oversight, algorithms risk entrenching and potentially worsening bias.

The way in which decisions are made, the potential biases which they are subject to and the impact these decisions have on individuals are highly context dependent. Our Review focuses on exploring bias in four key sectors: policing, financial services, recruitment and local government. These have been selected because they all involve significant decisions being made about individuals, there is evidence of the growing uptake of machine learning algorithms in the sectors and there is evidence of historic bias in decision-making within these sectors.

This Review seeks to answer three sets of questions:

1. **Data:** Do organisations and regulators have access to the data they require to adequately identify and mitigate bias?
2. **Tools and techniques:** What statistical and technical solutions are available now or will be required in future to identify and mitigate bias and which represent best practice?
3. **Governance:** Who should be responsible for governing, auditing and assuring these algorithmic decision-making systems?

Our work to date has led to some emerging insights that respond to these three sets of questions and will guide our subsequent work.

Data

While data itself is often the source of bias, it is also a core element of tackling the issue. It is common practice to avoid using data on protected characteristics (or proxies for those characteristics) as inputs into a decision-making process as to do so may be illegal. This is why, for example, organisations collecting diversity data on their employees ensure this is kept separate from decisions about employment and promotion.

However, some organisations do not collect diversity information at all, due to nervousness of a perception that this data might be used in a biased way. This then limits the ability to properly assess whether a system is leading to biased outcomes. For example, it would be impossible to establish the existence of a gender pay gap at a company without knowing whether each employee is a man or woman. This tension between the need to create algorithms which are blind to protected characteristics, while also checking for bias against those same characteristics, creates a challenge for organisations seeking to use data responsibly.

Tools and techniques

As the systems which inform decision-making become increasingly complex and data intensive, it can be difficult to establish if and where bias has originated.

Our early work suggests that new approaches to identifying and mitigating bias are required and we know that specific tools are already starting to be developed. This seems particularly true of sectors such as financial services which is highly-regulated, data-rich, and has a long history of using advanced models to make complex decisions such as the pricing of insurance rates. Some of these tools are being developed in-house, some are commercially available and others are being developed on an open-source basis.

However, there is limited understanding of the full range of tools and approaches available (current and potential) and what constitutes best practice. This makes it difficult for organisations that want to mitigate bias in their decision-making processes to know how to proceed and which tools and techniques they should use.

Governance

Data-gathering and analytical tools can help to identify the presence of bias in decision-making, but this is only a first step. We must subsequently make decisions that require value judgements and trade-offs between competing values. Humans are often trusted to make these trade-offs without having to explicitly state how much weight they have put on different considerations. Algorithms are different. They are programmed to make trade-offs according to unambiguous rules. This presents new challenges.

Furthermore, these tools must be used as part of a system of governance that is demonstrably trust-worthy; this may require new functions and actors – such as third party auditors – to independently verify claims made by organisations about how their algorithms operate.

We are drawing on a range of established governance principles[3] to inform this work. Effective human accountability for the use and performance of algorithmic tools will be critical regardless of context. However, the form of that accountability and the mechanisms required to make it effective will differ. Our sector approach will allow us to test this hypothesis, as well as explore what is required to operationalise ethical approaches in practice.

[3] www.gov.uk/government/publications/the-centre-for-data-ethics-and-innovations-approach-to-the-governance-of-data-driven-technology

1. Explaining the issue

Defining bias in the context of decision-making is challenging. In general usage, when we describe a decision as biased, what we mean is that it is not only skewed, but skewed in a way which is unfair. The decision has been made with reference to characteristics which are not justifiably relevant to the outcome, for example loan decisions that are systematically more favourable to men or women with otherwise similar financial situations.

Algorithms can be supportive of good decision-making, reduce human error and combat existing systemic biases. But issues can arise if, instead, algorithms begin to reinforce problematic biases, for example because of errors in design or because of biases in the underlying datasets. When these algorithms are then used to support important decisions about people's lives, for example determining whether they are invited to a job interview, they have the potential to cause serious harm.

The landscape summary[4] of the academic, policy and other literature relating to bias in algorithmic decision-making, commissioned by the CDEI, illustrates the complexities of this issue and highlights both the significant potential of these technologies to challenge biased decision-making and the risks that these same technologies could exacerbate existing biases. This Landscape Summary has informed our understanding and analysis of the issue.

Bias in algorithmic decision-making

Bias in decision-making is not new; it has long been a feature of human-led processes.

There are various points at which bias can enter human decision-making processes: the data or evidence may be biased; the process for assessing evidence may be problematic; or those taking subsequent actions may bring their own biases into the process.

Decision-making processes which are driven by algorithms, either entirely, or as a support to human decision-makers, share the same vulnerabilities: the input data, the design or performance of the algorithm itself, and the way in which its outputs are acted upon by humans, are all stages at which bias could be introduced, entrenched or amplified.

[4] www.gov.uk/government/publications/landscape-summaries-commissioned-by-the-centre-for-data-ethics-and-innovation

To understand bias in algorithmic decision-making, it is critical therefore to understand the multiple points at which bias can enter the decision-making process and the ways in which human and algorithmic biases interact.

Challenges in addressing bias

Defining fairness

Notions of 'fairness' are not universal, nor are they always consistent. Fairness can mean different things in different contexts and even different things within the same context. For example, procedural fairness is concerned with 'fair *treatment*' of people (*how* a decision is made) while outcome fairness prioritises 'fair *impact*' on people (*what* decisions are made). A 'fair' process may still produce 'unfair' results, and vice versa, depending on your perspective.

It is often impossible to optimise simultaneously for different definitions of fairness. This has been illustrated in the controversy surrounding the COMPAS algorithm, which has been used in courts in the United States.

COMPAS provides a risk score for the likelihood a defendant will reoffend and judges can use this assessment to inform decisions about whether to grant bail to a defendant awaiting trial. Some have argued the system is biased because a higher proportion of black defendants were put in the medium-high risk category but did not go on to reoffend compared to their white counterparts. Others have argued that the system *is* fair because defendants who have the same risk of reoffending go on to reoffend at the same rates, regardless of race. Both of these arguments are verifiably true, but they rely on different definitions of fairness; and it is mathematically impossible to produce results that satisfy both definitions at the same time.[5]

In human systems, it is possible to leave a degree of ambiguity about which definition is being applied. An algorithm, however, has to be unambiguous. It can be designed to meet either one of these definitions but it cannot meet all definitions of fairness at the same time. Humans must choose between them.

Direct and indirect bias

UK law seeks to protect people from discrimination on the basis of certain characteristics, such as their race or sex. The choice of these characteristics is a recognition that they have been used to unfairly discriminate in the past and that, as a society, we have deemed this discrimination unacceptable. For example, programming an algorithm to automatically reject female applicants for a job purely on the basis that they are women would be unlawful.

[5] www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/?utm_term=.df4a25bd4d61

Protected characteristics

The Equality Act 2010[6] makes it illegal to discriminate against someone on the basis of their:

- age
- disability
- gender reassignment
- marriage and civil partnership
- pregnancy and maternity
- race
- religion or belief
- sex
- sexual orientation

However, the use of data-driven technology may create or embed indirect biases by informing decisions that apply to everyone while having a disproportionate impact on some groups. For example, a credit-scoring algorithm may rate consumers who routinely buy clothes in certain kinds of shop less favourably because the algorithm indicates that this is a good predictor that they are less likely to pay back loans. However, if these shops are largely selling women's clothes, the algorithm will recommend fewer loans to women. The lawfulness of this type of decision-making is less clear and depends on judgements about the extent to which such selection methods are a proportionate means of achieving a legitimate aim.[7]

The increasing use of data and algorithms does not necessarily increase the risk of indirect bias in the world, although that is a possibility. However, the use of data and algorithms has the potential to increase our ability to identify indirect bias and our obligation to address the issue.

Decision-making, algorithmic or otherwise, can of course also be biased against characteristics which may not be protected in law, but which may be considered unfair, such as socio-ecomic background. In addition, the use of algorithms increases the chances of discrimination against characteristics that are not obvious or visible. For example, an algorithm might be effective at identifying people who lack financial literacy and use this to set interest rates or repayment terms.

[6] www.legislation.gov.uk/ukpga/2010/15/part/2/chapter/1

[7] www.equalityhumanrights.com/en/advice-and-guidance/commonly-used-terms-equal-rights

Mitigating bias

Blinding algorithms to demographic differences and proxies for these differences does not always lead to fairer outcomes. For example, preventing an algorithm designed to calculate the risk of criminals reoffending from taking into account their sex, would likely result in disproportionately harsher sentences for women overall as women tend to reoffend less often than men.[8] By excluding sex, the algorithm becomes less accurate for women and so, arguably, less fair.

Being blind to differences may also make it impossible to know whether an algorithm is being indirectly biased. For example, to protect against the risk of making discriminatory recruitment decisions, an organisation might seek to remove data that could identify the sex or ethnicity of job candidates from their decision-making process. That will ensure that these features cannot be considered directly by an algorithm designed to screen job applicants.

However, this approach does not remove the possibility of a machine learning algorithm using data that might be an effective proxy for these characteristics, for example, postcodes that correlate closely with race. In this example, the removal of data on ethnicity from the dataset may make it impossible to evaluate whether indirect bias is taking place. This highlights an important tension: to avoid '*disparate treatment'* as part of the decision-making process, sensitive attributes should not be considered by the algorithm. On the other hand, in order to assess '*disparate impact'*, sensitive attributes must be examined by those responsible to check if a given algorithm is fair.[9]

[8] www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/

[9] N. Kilbertus, A. Gascon, M. Kusner, M. Veale, K. P. Gummadi and A. Weller. Blind Justice: Fairness with Encrypted Sensitive Attributes. In the International Conference on Machine Learning (ICML), 2018

2. Scope of the Review

The previous section sets out the issue of bias in decision-making processes generally, explaining how data-driven technology has the potential to mitigate or exacerbate bias.

Our remit is to address bias that may be created, entrenched or aplified by algorithmic decision-making, rather than bias in decision-making more generally. In this context, we are looking particularly at:

- Bias which occurs against protected characteristics, although other groups which may not be formally protected by law will also be in scope; we are focusing on situations where the lawfulness of a decision-making process might be ambiguous or where illegal bias could be obscured by data or model complexity.
- Decision-making systems which have the potential to lead to individuals or groups being treated systematically unfairly; our work is exploring competing approaches to fairness and will expose the trade-offs involved when considering how we define what is fair or unfair.

Given the challenges of determining what is 'fair', it is important to understand the status quo prior to the introduction of data-driven technology in any given context. It is unlikely that any decision-making system will be completely free of potential bias, but it is possible that an algorithmic approach, depending on how it is engineered and checked, could be demonstrated to be significantly better than the system which predates it. This understanding of historical context is relevant to assess the level of potential harm or benefit to be derived from the new approach.

3. Our approach

This Review seeks to answer three sets of questions:

1. **Data:** Do organisations and regulators have access to the data they require to adequately identify and mitigate bias?
2. **Tools and techniques:** What statistical and technical solutions are available now or will be required in future to identify and mitigate bias and which represent best practice?
3. **Governance:** Who should be responsible for governing, auditing and assuring algorithmic decision-making systems?

The ethical questions in relation to bias in algorithmic decision-making vary depending on the context and sector. We have, therefore, selected four areas of focus to illustrate the range of issues and are exploring them in depth. These are policing, financial services, recruitment and local government.

All these sectors have the following in common:

- They involve making decisions at scale about individuals which may have significant impacts on those individuals' lives.
- There is a growing interest in the use of algorithmic decision-making tools in these sectors, in particular involving machine learning.
- There is evidence of historic bias in decision-making within these sectors, leading to risks of this being perpetuated by the introduction of algorithms.

We have chosen two public sector uses of algorithmic decision-making (policing and local government) and two predominantly private sector uses (financial services and recruitment). Investigating specific aspects of each of these sectors will provide us with evidence to consider how methods to mitigate bias can also be applied to other sectors more widely.

We are also drawing on work being done elsewhere. For example, the Information Commissioner's Office's (ICO) framework for auditing artifical intelligence is of particular interest as it includes work on how artifical intelligence can play a part in maintaining or amplifying human biases and discrimination.[10]

The methods we are using to answer these questions vary according to the sector and are described in the next section.

[10] https://ai-auditingframework.blogspot.com/

4. Progress to date

We are taking a phased approach to each sector, starting with policing, then moving onto financial services and recruitment, with our work on local government starting in autumn 2019. In this way we are able to refine our thinking as the Review progresses.

Evidence collection

We have spent the first stage of this Review collecting and analysing evidence. Earlier this year, we commissioned a team of academics led by Michael Rovatsos of the University of Edinburgh to conduct an assessment of the current academic, policy and other literature relating to bias in algorithmic decision-making.[11]

We will publish a summary of the wide-ranging responses we received to our Call for Evidence[12] asking individuals, groups and organisations to come forward with information on bias in algorithmic decision-making with a particular focus on the four sectors identified.

We have also been conducting ongoing stakeholder engagement with individuals, groups and organisations with an interest in algorithmic bias, including technology firms, trade bodies for relevant sectors, government departments, regulators, civil society organisations, and academics.

Alongside this, we continue to analyse reports and policy documents from the UK and internationally which relate to bias in algorithmic decision-making. We are also making use of additional legal and technical expertise as required.

Policing

Background

Our digital society is creating new and profound challenges for the criminal justice system. The volume of digital forensic material being seized for crimes is higher than ever, and the police are interacting with individuals used to living far more of their lives online.[13] This has led to the police holding more and more data, which they are under increasing pressure to manage more effectively in

[11] www.gov.uk/government/publications/landscape-summaries-commissioned-by-the-centre-for-data-ethics-and-innovation

[12] www.gov.uk/government/publications/the-centre-for-data-ethics-and-innovation-calls-for-evidence-on-online-targeting-and-bias-in-algorithmic-decision-making

[13] Ian Kearns and Rick Muir, 'Data Driven Policing and Public Value', The Police Foundation, p.2, www.police-foundation.org.uk/2017/wp-content/uploads/2010/10/data_driven_policing_final.pdf

order to identify connections and predict future risks. Given the rapidly changing digital context police face, there are significant opportunities for police to embrace new technologies and realise the benefits in terms of efficiencies and smarter use of data.

While a large amount of this data may not lend itself to sophisticated analysis, data such as crime location can provide police forces with a deeper understanding of crime patterns, allowing them to better target resources in those areas. For example, Avon and Somerset Police are employing software to collate data across different databases and visualise it in a way that helps police officers better understand a range of issues, including officer deployment.

Predictive tools may have the potential to improve the safety of citizens and it could therefore be argued that police forces have a social obligation to use them. Moreover, as existing manual risk assessment methods used by the police are challenged for being unfair and lacking rigour[14], more sophisticated tools could help remove certain biases inherent in an entirely human decision-making process and so provide a more objective assessment.

However, these tools carry significant ethical risks. As set out in section 1, one of the ways that bias can enter a decision-making process is if the underlying data or evidence carries bias within it. There is a possibility that the police data which would be required to train any machine learning approaches could be of variable quality and contain significant historic bias. Algorithms learning from biased or incomplete historic data in this way could perpetuate or exacerbate biased criminal justice outcomes for certain groups or individuals.

Focus of policing workstrand

This Review focuses on predictive analytics in policing, defined as ‘taking data from disparate sources, analysing them and then using results to anticipate, prevent and respond more effectively to future crime.’[15] To date, predictive analytic technology has been used in the following ways in policing:

1. **Predictive crime mapping**: Use of technology to predict geospatial locations where crime is likely to happen in the near future and preemptively deploy resources to where they are most needed.

2. **Algorithmic decision-support:** Use of algorithmic risk assessment tools to make predictions related to individuals, for instance to identify high-risk offenders whose past behaviour indicates they may be at increased risk of offending and reoffending in the near future.[16]

In the UK, predictive analytical tools have been used by the police for more than ten years in ‘predictive crime mapping’. However, the use of ‘algorithmic

[14] Alexander Babuta, Marion Oswald and Christine Rinik, ‘Machine Learning Algorithms and Police Decision-Making Legal, Ethical and Regulatory Challenges’, Royal United Services Institute, p.10, https://rusi.org/sites/default/files/201809_whr_3-18_machine_learning_algorithms.pdf.pdf

[15] Jennifer Backner, ‘Predictive Policing: Preventing Crime with Data and Analytics’, IBM Center for the Business of Government, Improving Performance Series, 2013, p.4

[16] Alexander Babuta, Marion Oswald and Christine Rinik, ‘Machine Learning Algorithms and Police Decision-Making Legal, Ethical and Regulatory Challenges’, Royal United Services Institute, p.3

decision-support' in policing is in its infancy and the potential outcomes and indirect consequences of these tools are still poorly understood.

There are many other uses of data-driven technology for law enforcement purposes. For example, facial recognition technology (FRT) has received significant media and public attention over the last year, with two legal challenges launched against South Wales and London Metropolitan Police. Whilst this has brought to light important ethical concerns around the use of FRT, the issues associated with this technology go wider than bias. The CDEI is planning to produce a briefing paper on FRT later in the autumn which will examine these wider ethical concerns and will not be limited to the use of FRT by the police.

We appreciate that data-driven technology is also being used more widely across the criminal justice system, for example in assessing an individual's risk of reoffending and risk of harm to others. To some extent, the tools in use present similar ethical dilemmas to those associated with predictive analytics in policing and we are liaising with the Ministry of Justice on how our work may be complementary.

Our approach

Given the potentially significant impact that data-driven technology can have on citizens' lives, we believe that new technologies should be trialled in a controlled way prior to implementation, to establish whether or not a certain tool is likely to improve the effectiveness of a policing function. However, there is currently no clear framework for how the police should conduct such trials and deploy this technology. This creates the risk that technologies will be adopted at scale without proper consideration of their potential to generate biased outputs or of how to address the wider ethical concerns. At the same time, public scrutiny of the tools is growing.[17]

In our conversations with police representatives and other key stakeholders, including think-tanks and ethics bodies, we have heard active calls for centralised policy guidance. This has been echoed by police leaders such as the Commissioner of the Metropolitan Police, Cressida Dick, who has publicly noted[18] the lack of clear guidance on the use of machine learning algorithms in policing. The risk has also been flagged in Parliament by the All Party Parliamentary Group on Data Analytics, which noted that predictive analytical tools in policing represent the most controversial area considered by their research.[19] Moreover, The Law Society's recent report[20] noted the clear lack of explicit standards, best practice, and openness or transparency about the use of algorithmic systems in the criminal justice system across England and Wales and the need to develop a range of new mechanisms to improve oversight.

[17] www.bbc.co.uk/news/stories-48718948

[18] www.youtube.com/watch?v=KRuetVfovnI

[19] www.policyconnect.org.uk/appgda/research/trust-transparency-and-technology-building-data-policies-public-good

[20] www.lawsociety.org.uk/support-services/research-trends/algorithm-use-in-the-criminal-justice-system-report/

The Home Office is aware of the ethical challenges of predictive analytics and has recently expanded the mandate of the Biometrics and Forensics Ethics Group[21], a non-departmental public body, to include ethical issues relating to large, complex datasets. Moreover, individual police forces are bolstering their own ethical scrutiny of the predictive analytics projects they are developing, as demonstrated by the setting up of the West Midlands Police and Crime Commissioner's Ethics Committee.[22]

To ensure these tools receive appropriate guidance and oversight, the CDEI is working collaboratively with the Home Office, policing sector bodies (including individual forces), the technology sector and civil society to develop a Code of Practice to support police forces to develop, trial and deploy predictive analytical tools more effectively. The Code of Practice will look at the oversight and accountability that is needed to govern this technology effectively.

We have established the following partnerships to inform our work:

Commissioned research by the Royal United Services Institute (RUSI): We have commissioned the think-tank RUSI to undertake research, including interviews and roundtables with police forces, academics, civil society, and policymakers, to underpin the development of the Code of Practice. RUSI will produce an interim briefing with initial research findings in September 2019 and a final report in early 2020.

Partnership with the Cabinet Office's Race Disparity Unit (RDU): We are working with the RDU, a UK Government unit which collates, analyses and publishes government data on the experiences of people from different ethnic backgrounds in order to drive policy change where disparities are found. We are drawing on its expertise to better understand how algorithmic decision-making could disproportionately impact ethnic minorities. The RDU is feeding into our roundtables with RUSI and supporting our partnerships with individual police forces.

Partnership with West Midlands Police: We will co-develop and test our draft Code of Practice with West Midlands Police in order to ensure it is useful and meets the required needs of the police. While we will also work with other police forces to develop the Code of Practice, West Midlands Police, as mentioned above, already has an in-house data analytics team currently developing predictive analytical tools and an independent ethics panel run out of the Office of the West Midlands Police and Crime Commissioner and are therefore well-positioned to work with us.[23]

[21] www.gov.uk/government/news/ethics-group-to-oversee-use-of-large-data-sets-by-the-home-office

[22] www.westmidlands-pcc.gov.uk/ethics-committee/

[23] https://techswitchcf.com/2019/02/09/west-midlands-police-turns-to-predictive-analytics/

Financial Services

Background

In making decisions to price and grant credit and insurance policies to individuals, financial services firms exert considerable influence over individuals' lives. As a sector which is already used to employing advanced mathematics to guide decision-making in a highly-regulated environment, finance is well-placed to embrace the most advanced data-driven technology to make better decisions about which products to offer to which customers. However, the history of finance includes profoundly discriminatory practices, for example, redlining majority ethnic minority postcodes[24] and requiring a woman's husband's signature on a mortgage.[25] While the most egregious of these no longer happen, these organisations still operate in a socio-economic environment where financial resources are not spread evenly between different groups. As a result these embedded historical biases may be reflected in the data held.

The financial services sector relies on being able to make decisions at scale about individuals' financial futures based on predictions of likely behaviour, for example, in relation to repaying debts. Broadly speaking, the success of their business models are often based on how accurately they can make these predictions. The financial services sector is exploring the disruptive change that will come from incorporating larger datasets, new sources of data (such as data from social media profiles), and more sophisticated machine learning into its decision-making processes.

Using more data and better algorithms may yield better risk prediction and mean fewer people are denied loans because of inaccurate credit scoring. It may also enable population groups who have historically found it difficult to access credit (because of a paucity of data about them from traditional sources) to gain better access in future. At the same time, more data and more complex algorithms increases the potential for the introduction of indirect bias via proxy as well as the ability to detect and mitigate it.

Focus of financial services workstrand

Our focus in this workstrand is on credit and insurance decisions taken about individual customers. The financial services sector has a long established history of using statistical methods, for example, credit rating[26] in highly-regulated markets. As such, we are focusing on recent technological developments, in particular the use of data from non-traditional sources, such as social media, and emerging machine learning approaches.

[24] https://ncrc.org/holc/

[25] www.moneysupermarket.com/credit-cards/womens-financial-rights/

[26] www.investopedia.com/articles/bonds/09/history-credit-rating-agencies.asp

Our approach

Drawing on our own research, the landscape summary, and our recent Call for Evidence, we are carrying out structured interviews with key stakeholders in financial services to identify the main barriers faced in identifying and mitigating bias. We then plan to conduct a survey of algorithmic bias identification tools currently available and assess the strengths and weaknesses of these approaches to begin to establish best practice standards. We will also consider how tools may need to develop to deal with the application of emerging data-driven technology, including the use of machine learning algorithms. Finally, we will work with stakeholders to identify potential governance arrangements to oversee the mitigation of bias across the financial services sector.

In addition, we have commissioned the Behavioural Insights Team (BIT)[27] to undertake research into public perceptions of fairness in the context of credit rating. BIT is running an experiment on its Predictiv platform which will help us understand how fair the public feel it is if a bank uses an algorithm that produces different outcomes for men and women, or different ethnic groups, and where the justification for this is unclear. The experiment will test how this affects participants' own financial decision-making and which banks they request loans from. By working with BIT in an experimental format, we can begin to understand more about trade-offs between different kinds of fairness, how the public perceive these and how these can be incorporated into best practice. This will inform how we work with the industry to develop governance frameworks making sure these reflect our latest understanding of public perceptions of fairness in the context of access to credit.

Recruitment

Background

The decision to shortlist, interview and employ someone in a particular job can have a profound influence on that individual's life. On a societal level, the systemic exclusion or over-representation of certain groups in certain professions can embed social inequalities. While still relatively nascent, the use of data-driven technology in recruitment is predicted to be a growing trend over the next few years.[28] These technologies could range from relatively simple text scanning technology, to more complex content analysis and even artificial intelligence-led interviews.[29] Historically, particular jobs have not been equally accessible to all and current initiatives such as gender pay gap reporting demonstrate that these biases continue to manifest themselves in our present-day work structures. Incidents such as Amazon withdrawing its recruitment

[27] www.bi.team/

[28] www.forbes.com/sites/forbescoachescouncil/2018/08/10/10-ways-artificial-intelligence-will-change-recruitment-practices/#22e399d53a2c

[29] https://towardsdatascience.com/your-next-job-interview-may-be-with-an-ai-robot-34dbf4da6340

algorithm because of gender bias[30] suggest that algorithmic technology, if not designed carefully, has the potential to further embed these biases.

The potential for intelligent algorithmic systems to improve current job matching services is significant. Being able to recommend jobs to people that they might not search for or think themselves able to apply for is a development that many in the recruitment world are now exploring. Encouraging the development of these systems could benefit many people but recruiters will need to ensure that their recommendations are not discriminatory.

Focus of recruitment workstrand

Our focus is on the use of algorithms to automate (or partially automate) recruitment decisions. This can range from bulk screening of CVs and applications, to providing recommendations to human decision-makers, for example, on who to invite to interview, to using artificial intelligence to analyse candidates' performance in interviews.

Our approach

Drawing on our own research, the landscape summary, and our recent Call for Evidence, we are carrying out structured interviews with key stakeholders in recruitment. We aim to ascertain the key barriers faced in identifying and mitigating bias when using data-driven technology to support recruitment decisions. We then plan to conduct a survey of current algorithmic bias identification tools and assess the strengths and weaknesses of these approaches, to begin to establish best practice standards.

Vendors of algorithmic recruitment tools, such as employment assessments to screen candidates, are exploring bias mitigation approaches but lack clear guidance on how to develop these.[31] Our work will survey existing tools and practices, supporting companies to understand the range of possible bias-mitigation approaches available and to shape industry standards and best practice.

We appreciate the sector has no clear regulator. As with financial services, we will consider with stakeholders the potential governance arrangements for overseeing the mitigation of bias across this sector.

Local Government

Background

Local authorities are responsible for making significant decisions about individuals on a daily basis, in particular, decisions about children judged to be at risk and in need of support or intervention. The individuals making these decisions are often working under significant time and resource pressures.

[30] www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

[31] Manish Raghavan, Solon Barocas, Jon Kleinberg, Karen Levy, 'Mitigating Bias in Algorithmic Employment Screening: Evaluating Claims and Practices' – https://arxiv.org/abs/1906.09208

They are required to draw on complex sources of evidence as well as their professional judgement.

Academic and commercial developers are starting to produce predictive analytical tools aimed at this sector,[32] and some local authorities are starting to consider their potential. Relatively limited research is available in this area, but given the vulnerable communities involved and concerns over the quality of some local government data, there are risks around whether the use of data-driven technology could exacerbate the risk of bias. Mitigating this early will be critical to ensuring that the tools serve their intended function of supporting improved decision-making.

Focus of local government workstrand

Our focus is on the use of data-driven technology to guide individuals working in local government who are making decisions about the individuals and communities they serve. In particular, we are interested in the use of these tools to support social care decisions to target interventions at children judged to be at risk. This is an area where the stakes for individuals are particularly high. It is also the area where predictive analytics is currently most frequently applied in local government, albeit in trial or exploratory form, in particular to assist social workers in deciding whether to refer cases for further action or not.[33]

Our approach

The CDEI will formally begin work on this sector the autumn. So far, we have spoken with a selection of local authorities, academics and technology providers and received formal input via our Call for Evidence. We are still in the process of defining the approach we will take to this sector and we would welcome further input from interested stakeholders.

[32] Predictive Analytics, https://troubledfamilies.blog.gov.uk/2018/05/14/predictive-analytics/

[33] https://smartcities.oii.ox.ac.uk/wp-content/uploads/sites/64/2019/04/Data-Science-for-Local-Government.pdf

5. Emerging insights

This Review seeks to answer three sets of questions:

1. **Data:** Do organisations and regulators have access to the data they require to adequately identify and mitigate bias?
2. **Tools and techniques:** What statistical and technical solutions are available now and in the future to identify and mitigate bias and which of these represent best practice?
3. **Governance:** Who should be responsible for governing, auditing and assuring algorithmic decision-making systems?

Our work so far has led to some emerging insights about each of these areas and our sector focus is helping us to understand how these questions play out across different contexts. This will guide our subsequent work.

Data: Data is fundamental to training decision-making algorithms and evaluating them for possible bias. Data itself is often the source of bias but, at the same time, it is a core element of tackling the issue. Datasets frequently contain significant biases, either because they are incomplete and unrepresentative or because they are accurately reflecting historic patterns of bias. For example, our early research on the use of predictive analytical tools in policing suggests that one of the key issues with regards to the use of this technology is potential bias embedded in historic datasets. As datasets become more complex and our ability to analyse them more sophisticated, the risk of bias is likely to increase as new proxies and patterns of behaviour emerge in the datasets. Algorithms may then be used to guide decisions in ways which society may judge to be unfair.

It is common practice to avoid using data on protected characteristics (or proxies for those characteristics) as inputs into decision-making algorithms, as to do so is likely to be illegally discriminatory. However, understanding the distribution of protected characteristics among the individuals affected by a decision is necessary to identify biased impact. For example, it is impossible to establish the existence of a gender pay gap at a company without knowing whether each employee is a man or woman. This tension between the need to create algorithms which are blind to protected characteristics while also checking for bias against those same characteristics creates a challenge for organisations seeking to use data responsibly.

Tools and techniques: As the systems which inform decision-making become increasingly complex and data intensive, it can be difficult to establish where bias has originated. Organisations using decision-making algorithms have an interest in evaluating potential unintended biases emerging from these systems, creating a need for bias identification tools and techniques.

In response to this need, approaches to evaluating decision-making algorithms for bias are beginning to be proposed, either by academics in literature or by interested groups or companies as products or services. There is limited understanding of the full range of these approaches. Some are freely available, while others are commercially marketed. Our sector approach suggests that certain sectors, for example, financial services, are more advanced in their thinking on this. However, across the board there appears to be a lack of clarity over the relative strengths and weaknesses of these tools. Organisations are also pursuing in-house approaches to bias identification but it can be hard for organisations to know how they compare to other tools available. This limited information makes it difficult for those that want to follow best practice to evaluate their processes for possible bias.

Governance: Algorithms are usually a component in a broader decision-making process involving human decision-makers. It is critical that governance approaches cover this broader context and do not focus exclusively on the algorithmic tools themselves.

Data gathering and analytical tools can help to understand the presence of bias in decision-making, but this is only a first step. We must subsequently decide how far to mitigate bias and how we should govern our approach to doing so. These decisions require value judgements and trade-offs between competing values. Humans are often trusted to make these trade-offs without having to explicitly state how much weight they have put on different considerations. Algorithms are different. They are programmed to make trade-offs according to rules and their decisions can be interrogated and made explicit. This requires a different approach to accountability.

Decision-makers are likely to face significant trade-offs, for example, between different kinds of fairness and between fairness and accuracy. Knowing what standards systems should be expected to operate to is difficult, in particular, whether it is sufficient that they are less biased than equivalent human systems or whether they should be held to higher standards.

There is currently limited guidance and a lack of consensus about how to make these choices or even how to have constructive and open conversations about them. These choices are likely to be highly context specific and as such, the way they are made, governed and audited will need to be considered on a sector by sector basis. For example, in the policing sector we are developing a Code of Practice in collaboration with the sector, but this is unlikely to be an appropriate approach for financial services or recruitment given their different operating and regulatory environments.

A certain level of transparency about the performance of algorithms will be necessary for customers and citizens to be able to trust that they are fair. Giving developers of algorithms space and opportunity to test algorithms against standard datasets or to benchmark performance against industry standards may enable the development of a consensus about the appropriate definitions of fairness. New functions and actors, such as third party auditors, may also be required to independently verify claims made by organisations about how their algorithms operate.

6. Next steps

As we progress our work in each sector, we will continue to explore our questions around data, tools and techniques, and governance. Next steps over the next eight months include:

Call for Evidence: we will publish a summary of responses received later in the summer.

Policing: we will publish a draft Code of Practice in the autumn for consultation and then finalise the Code of Practice in early 2020.

Financial Services: we will continue to engage with sector stakeholders and commission research to inform our final recommendations.

Recruitment: we will continue to engage with sector stakeholders and commission research to inform our final recommendations.

Local Government: we will share more details of our planned approach to this work in the autumn.

We will submit a final report with recommendations to the Government in March 2020.