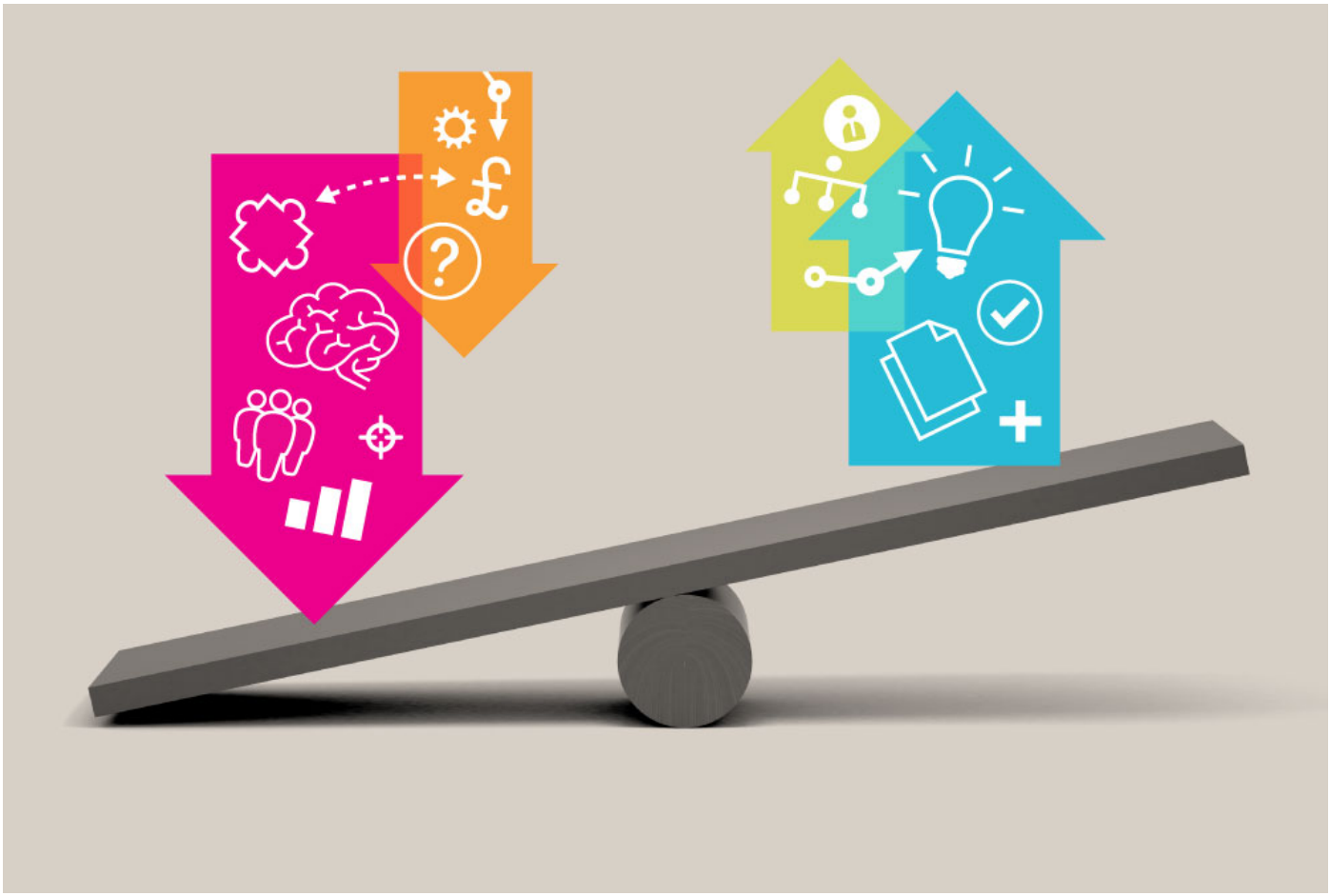
Human bias and discrimination in AI systems



As part of our AI auditing framework blog series, Reuben Binns, our Research Fellow in Artificial Intelligence (AI), and Valeria Gallo, Technology Policy adviser, look at how AI can play a part in maintaining or amplifying human biases and discrimination.

25 June 2019

This post is part of our ongoing Call for Input on developing the ICO framework for auditing AI. We encourage you to share your views by emailing us at AIAuditingFramework@ico.org.uk.

The fact that AI systems learn from data does not guarantee that their outputs will be free of human bias or discrimination. The data used to train and test AI systems, as well as the way they are designed, and used, are all factors that may lead AI systems to treat people less favourably, or put them at a relative disadvantage, on the basis of protected characteristics [1].

The UK anti-discrimination legislative framework, notably through the UK Equality Act 2010, offers individuals protection from discrimination, whether generated by a human or automated decision-making system.

The General Data Protection Regulation (GDPR) complements this framework, by introducing provisions that are specifically designed to protect data subjects' 'fundamental rights and freedoms' as a result of the processing of their personal data, including the right to non-discrimination. GDPR specifically notes that data controllers should take measures to prevent 'discriminatory effects on natural persons'.

In this post we explore what these provisions mean, in practice, in the context of AI. We will focus on how machine learning (ML) systems used to classify or make a prediction about individuals may lead to discrimination and we will explore some of the technical and organisational measures that can be adopted to manage this risk.

Why might an ML system lead to discrimination?

Let's take a hypothetical scenario:

A bank has developed a ML system to calculate the credit risk of potential customers. The bank will use the ML system to approve or reject loan applications. To train the system the bank has collected a large set of data containing a range of information about previous borrowers, such their occupation, income, age, and whether or not they repaid their loan. During testing, the bank wants to check against any possible gender bias, and finds the ML system is giving women lower credit scores, which would lead to

fewer loans being approved.

There are two main reasons why this might be:

1. **Imbalanced training data** The proportion the male vs. female sub-populations in the training data may not be balanced. For example, the training data may include a greater proportion of male borrowers because in the past fewer women applied for loans and therefore the bank doesn't have enough data about women.

   The ML algorithm will generate a statistical model designed to be the best fit for the data it is trained and tested on. If the male population is over-represented in the training data, the model will pay more attention to the statistical relationships that predict repayment rates for men, and less to any different statistical patterns that predict repayment rates for women.

   Put another way, because they are statistically less important, the model could systematically predict lower loan repayment rates for women, even if females in the training dataset are on average more likely to repay their loans than men.

   These issues will apply to any sub-population under-represented in the training data. For example, if a facial recognition model is trained on a disproportionate number of faces belonging to a particular ethnicity and gender (eg white males), it will perform better when recognising individuals in that group.

2. **Training data reflects past discrimination**

   The training data the model is based on may reflect past discrimination. For instance, if in the past women's loan applications were rejected more frequently than men's on the basis of gender, then any model

based on such training data is likely to reproduce the same pattern of discrimination.

Certain domains where discrimination has historically been a significant problem, such as policing or recruitment, are more likely to experience this problem.These issues can occur even if the training data does not contain any protected characteristics like gender or race. Several features in training data are often closely correlated with protected characteristics, eg occupation. These 'proxy variables' enable the model to reproduce patterns of discrimination associated with those characteristics, even if its designers did not intend this.

These problems can occur in any statistical model, but they are more likely to occur in ML systems because they can include a much greater number of variables. ML is more powerful than traditional statistical approaches because it is better at uncovering hidden patterns in data. However, these also include patterns that reflect discrimination.

Technical approaches to mitigate discrimination risk in ML models

There are various available approaches to deal with issues arising from training data. In cases of **imbalanced training data** it may be possible to balance it out by adding or removing data about under/overrepresented subsets of the population (eg adding more data about female loan applicants or removing data about men).

Alternatively, an organisation could train separate models, for example one for men and another for women, and design them to perform as well as possible on each sub-group. However, in some cases, creating different models for different protected classes could itself be a violation of non-discrimination law (eg different car insurance premiums for men and women).

In cases where the training **data reflects past discrimination**, organisations could either modify the data, change the learning process, or modify the model after training.

To support these approaches, computer scientists and applied statisticians have been developing different mathematical techniques to understand how ML models treat individuals from different groups and any discriminatory effects they may have on the individuals belonging to them. Data scientists often refer to this as algorithmic "fairness".

These approaches can be grouped in three broad categories:

1. **'Anti-classification'** – according to which a model is fair if it excludes protected characteristics from consideration. Some anti-classification approaches also try to identify and exclude proxies for protected characteristics (eg attendance at a single-sex school). This can be impractical as removing all possible proxies may leave very few predictively useful features. Also, it is often hard to know whether a particular variable is a proxy for a protected characteristic without further data collection and analysis.
2. **Outcome and error parity**, which compares how members of different protected groups [2] are treated by the model.
    - *Outcome parity*: a model is fair if it gives equal numbers of positive or negative outcomes to different groups.
    - *Error parity*: a model is fair if it gives equal numbers of errors to different groups. Error parity can be broken down into parity of false positives or false negatives (see our Accuracy blog post for more details).
3. **Equal calibration** – Calibration measures how closely the model's estimation of the likelihood of something happening matches the actual frequency of the event happening. According to 'equal calibration' a model is fair if it is equally calibrated between members of different protected groups. For instance, of those loan applicants who are

predicted to have a 90% chance of repayment, results should show an equal proportion of male and female applicants actually repaying. (NB: Equal calibration does not necessarily require good calibration, only that any imperfections should affect protected groups equally).

Unfortunately, these different measures are sometimes incompatible with each other, and therefore any conflicts will have to be considered carefully before selecting any particular approach(es). For example:

- Equal calibration is incompatible with false positive parity, unless there is an exactly equal number of people from different protected groups in each class.
- Attempting to achieve outcome parity while removing protected characteristics, as required by anti-classification measure, may result in the model finding and using irrelevant proxies in order to equalise outcomes.

These are only some of the potential technical approaches to understanding and mitigating bias and discrimination in ML systems, and organisations may choose or devise others.

What can organisations do?

The most appropriate approach to managing the risk of discriminatory outcomes in ML systems will depend on the particular domain, social and political context in which the organisation deploying the AI solution will operate.

Organisations should determine and document their approach to bias and discrimination mitigation from the very beginning of any AI application lifecycle, so that the appropriate safeguards and technical measures can be taken into account and put in place during the design and build phase.

Establishing clear policies and good practices for the procurement of high-

quality training and test data will be important, especially if organisations do not have enough data internally or have reason to believe it may be unbalanced or contain bias. Whether procured internally or externally, organisations should satisfy themselves that the data is representative of the population the ML system will be applied to. For example, for a high street bank operating in England and Wales, the training data could be compared against the most recent Census.

The organisation's governing body will be responsible for signing-off on the chosen approach to manage discrimination risk and is accountable for its compliance with data protection law. While they will be able to leverage expertise from technology leads and other internal or external subject matter experts, to be accountable board members will still need to have a sufficient understanding of the limitations and advantages of the different approaches. This will also be true for Data Protection Officers and senior staff in oversight functions, as they will be expected to provide ongoing advice and guidance on the appropriateness of any measure and safeguards put in place to mitigate discrimination risk.

Processing of personal data using ML systems is likely to trigger the requirement to carry out a Data Protection Impact Assessment (DPIA), for instance if they are used to carry out profiling on a large scale. As part of their DPIA, depending on the severity of the risks associated with an ML system and the ability to manage any potential discrimination risks, they may also be required to consult with data subjects or their representatives to seek their views.

In many cases, choosing between different risk management approaches will require trade-offs, including between safeguards for different protected characteristics and groups. These will need to be fully documented and signed-off on. Trade-offs driven by technical approaches will not always be obvious to non-technical staff so data scientists should highlight and explain these proactively to business owners, as well as to staff with responsibility

for risk management and data protection compliance. Technical leads should also be proactive in seeking domain-specific knowledge, including known proxies for protected characteristics, to inform algorithmic “fairness” approaches.

Organisations should undertake robust testing of any anti-discrimination measures, and should monitor the ML system’s performance on an ongoing basis. Risk management policies should clearly set out the process, and the person responsible, for the final validation of an ML system before deployment, or after an update.

For monitoring purposes, organisational policies should set out any variance tolerances against the selected Key Performance Metrics, as well as escalation and variance investigation procedures. Variance limits above which the ML system should stop being used should also be clearly set. If the organisation is replacing traditional decision-making systems with AI, they should consider running them concurrently for a period of time, and investigate any significant difference in the type of decisions (eg loan acceptance or rejection) for different protected groups between the two systems.

While it is not a legal requirement under data protection regulation, a diverse workforce is a powerful tool in identifying and managing bias and discrimination in AI systems, and in the organisation more generally.

Finally, this is an areas where best practice and technical approaches continue to develop. Organisations should invest the time and resources to ensure they continue to follow best practice and their staff remain appropriately trained on an ongoing basis. In some cases AI may actually provide an opportunity to uncover and address existing discrimination in traditional decision-making processes, and allow organisations to address any underlying discriminatory practices.

Broader considerations

Discussions on the risk of discrimination in AI systems feed into a much broader debate of the ethical and societal impact of AI. These are important discussions that the ICO is actively contributing to. For the purpose of our AI auditing framework however, our focus is on the reasonable steps we will expect organisations to take to demonstrate compliance with the existing data protection requirements.

In addition, the provision of the GDPR are only one part of the broader anti-discrimination regulatory framework, and it feels important to stress that data protection compliance alone may not be sufficient to satisfy additional regulatory requirements outside the ICO regulatory perimeter.

Your feedback

As usual, we would like to hear your views on this topic and genuinely welcome any feedback on our current thinking on the topic of discrimination in AI systems. In particular, we would appreciate your insights on the following two questions:

- If your organisation is already applying measures to detect and prevent discrimination in AI, what measures are you using or have you considered using?
- In some cases, if an organisation wishes to test the performance of their ML model on different protected groups, it may need access to test data containing labels for protected characteristics. In these cases, what are the best practices for balancing non-discrimination and privacy requirements?

We encourage you to share your views by emailing us at AIAuditingFramework@ico.org.uk.

Dr Reuben Binns, a researcher working on AI and data protection, joined the ICO on a fixed term fellowship in December 2018. During his two-year term, Dr Binns will research and investigate a framework for auditing algorithms and conduct further in-depth research activities in AI and machine learning.

Valeria Gallo is currently seconded to the ICO as a Technology Policy Adviser. She works with Reuben Binns, our Artificial Intelligence (AI) Research Fellow, on the development of the ICO Auditing Framework for AI. Prior to her secondment, Valeria was responsible for analysing and developing thought leadership on the impact of technological innovation on regulation and supervision of financial services firms.

Footnotes

[1] Sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation.

[2] Protected groups are identified in the Equality Act 2010 as group of

persons defined by reference to a particular characteristic against which is it illegal to discriminate.

Next blog