The Alan Turing Institute

Response to the Centre for Data Ethics and Innovation's review on Bias in Algorithmic Decision-Making

Response of The Alan Turing Institute to the Centre for Data Ethics and Innovation's review on Bias in Algorithmic Decision-Making

Introductory comments

The Alan Turing Institute welcomes the focus of the review on particular sectors. This reflects an understanding that the impact of bias on individuals affected by algorithmic decisions differs by context. It should nonetheless be remembered that bias in algorithmic decision-making may be a combination of biases in data sets, human decision-making in context, and machine learning algorithms, which can be shared, reinforced and amplified through cross-sector and government collaborations. Furthermore, technology companies operate across sectors. This debate must therefore examine cross-sectoral approaches in considering the question of bias.

The response below provides personal reflections from a number of researchers in The Alan Turing Institute's community on some of the questions in the review, touching on various sectors of focus, including consumer-targeted credit risk and credit scoring.

Responses to Questions

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

There is a risk to individual rights and freedoms in systems where algorithms determine who may get a mortgage, how much an insurance policy costs, and whether certain behaviours or places people visit can be flagged as 'suspicious' activities. For example, if someone does not have an algorithmic profile, they may be excluded from systems where diverse voices are needed for decision-making. However, to have such a profile can mean we are further monitored and tracked to make algorithmic predictions more accurate.

Indeed, one must question why the everyday life experience of a person, online and offline, should be used for behaviour prediction by default. Before the "age of data", data collection tended to be purpose-specific in the context of a particular business interest, such as marketing. We now have wide-ranging and pervasive data collection, and the re-purposing and re-selling of this data.

A key concern is imposing data-driven systems and automated decision-making mechanisms on our everyday lives without public understanding and consensus. Decision-making mechanisms might not include diverse views, due to lack of representation or training with

respect to existing societal inequalities. This makes it problematic to implement such systems widely and to expect everyone to experience the same levels of benefit and risk. These systems have the potential to identify, sort and categorise citizens, which can lead to further divides in society and discriminatory practices, including being denied access to certain services. There may, however, be compensating benefits due to incentives nudging people towards better behaviour – e.g. if people who go to the gym more are offered lower insurance premiums, then this may lead to people becoming healthier; although many might not be comfortable with this type of monitoring.

The inferences, and hence the decisions, made by algorithmic processes are another cause of concern. In crime prevention, an algorithm can flag certain activities as 'suspicious', which could for example include carrying two mobile phones with different SIM cards at the same time, or taking the SIM card out of the phone in certain locations and during certain hours. Once these activities are detected as 'anomalies' in a system, it is not clear who then gets to decide whether these could constitute grounds for further monitoring, surveillance and control.

In financial services, credit scoring relies on statistical models to estimate the probability that a debtor will repay a credit obligation. Its use in banks and regulated financial institutions (FIs) is normed by banking regulators, and the process is understood by all actors.[1] Transparency regulations in banking force FIs to train these models using interpretable functions, such as logistic regression or survival analysis. New actors, such as fintech companies, may be able to get an edge over regulated institutions if they are able to tap into other types of non-structured data.[2] Using mixed sources of data can be economically efficient for those firms able to use them,[3] for example app-based lenders. Some companies are already using mobile phone data, psychometrics, behavioural data inferred from phones, and social network data.

As with other fields, the ethical challenges in this situation are how to ensure fairness in lending whilst allowing for the appropriate use of these data sources. Currently, it is much easier for regulators to detect if, for example, gender is being inadvertently added to a model through the use of a variable highly correlated with gender, but when using complex,

1 Thomas, L. C., Crook, J., Edelman, D. (2017) Credit Scoring and Its Applications, Second Edition. SIAM, p380.

2 Such as images, text, audio, social network information (both internet-based and inferred from connections), location-based data and a long list of data sources currently being used in other fields.

3 Óskarsdóttir, M., Bravo, C., Sarraute, C., Vanthienen, J., & Baesens, B. (2019). The value of big data for credit scoring: Enhancing financial inclusion using mobile phone data and social network analytics. Applied Soft Computing, vol 74, pp26-39.

unstructured data sources this is more challenging. However, potential societal gains brought by greater access to credit for those who are able to repay, incentivise finding a useful compromise.

2.1 To what extent (either currently or in the future) do we know whether algorithmic decision-making is subject to bias?

Bias can arise from the lack of transparency in black-box methods, and the modelling done by data scientists which feeds into machine-learning methods. Such modelling also requires transparency, as some data scientists might embed bias or not be sufficiently conscious of potential biases. Moreover, the data itself is a significant potential source of bias, and more of a focus needs to be put on its quality and limitations. In recruitment, for instance, machine learning is typically used to train a system to reproduce past decisions, which were often not made fairly. Some checks can be automated, such as verifying whether minority groups are underrepresented in the training data, but selection biases are not always evident from the information available. This may require access to baseline group distribution data that is not readily available, and correcting for past biased selection processes is not always clear if we do not know how they took place. Nevertheless, developers of decision-making tools should strive to document, to the best of their knowledge, all bias issues in their data and the assumptions already made to correct for biases – before they apply any machine learning.

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.

The risk is very high in the model training phase. If corrective measures are not taken, biases can be ingrained in the weights of the model and thus perpetuated for as long as the model is in use. This could lead to prediction bias which can only be detected if it is being actively investigated.

Human bias can also 'leak in' (explicitly, implicitly, or accidentally) through target functions, which set the criteria for which the algorithm is instructed to optimise. For example, if HMRC is trying to target an investigation into tax evasion, it could instruct the algorithm to maximise the number of people found (i.e. send inspectors to taxi drivers) or the amount of money recovered (i.e. send its inspectors to city bankers). The choice of success metric, which embeds assumptions about the relative costs of various types of error, such as false positives and false negatives, is central to social justice.

It is implicitly assumed that we are mainly concerned about existing human bias finding its way into artificial intelligence. While this concern is legitimate, we must also be aware that mathematically optimal rules are not necessarily socially acceptable. Deciding loan interest

rates based on postcode, social class, or gender may be good mathematical predictors of risk, but this does not mean they are socially just.

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?

If the provenance of data used for training algorithms is not known or trustworthy, definitions are arising in the research community to force predictions to respect categorical constraints. For example, predictions should not be sensitive to gender, race, and so on. More generally, prediction shouldn't be sensitive to so-called protected attributes (as well as proxy attributes that might indirectly reveal the protected attribute).

However, there remain four major challenges:

1. Some of these definitions yield counterintuitive outcomes, and correcting for such an outcome can yield a definition that is not compatible with the other. For example, when implementing so-called demographic parity, which says that prediction probabilities should remain the same when, say, changing the gender, it works well on hiring data in the sense that both men and women would be hired with equal likelihood, all other things being equal, but when applied to predicting the likelihood of committing a crime, women, who form less than 10% of the incarcerated population in the UK, will be just as likely to be considered high risk individuals as men.[4]
2. It is unclear if there is a uniform way to implement such definitions, especially in a scalable manner.[5]
3. Identifying protected and proxy attributes can be very difficult, and if the data is sparse, labelling too many useful attributes as being protected/proxy could lead to uninteresting predictions.
4. Once bias has been identified, it is unclear what the next step is. The research community has suggested the notion of attribute correction, which attempts to rectify and adjust, say, the salary distribution between men and women if it is seen that men are paid more on average for the same skills.[6] This only addresses the symptoms and not the cause, so a deeper qualitative analysis is needed to understand how bias can be rectified at a societal

[4] A follow-up definition called equality of opportunity raises other problems. See, for example, discussions in https://arxiv.org/abs/1905.07026.

[5] This has prompted implementation strategies such as the one in the paper linked to in footnote 4.

[6] A more granular notion is put forward in some recent papers, including the link in footnote 4, that attempts percentile-based analysis and correction.

level. It will also be important to consider what other features could be important, yet may not currently be collected – for example, number of hours worked or the risk of injury.

A further issue is that regulators and scientists largely do not know what most commercial users of data-driven algorithmic tools are doing, and have no way to access this. Without this understanding of commercial activity, any attempts to enforce rules to reduce bias will have little impact.

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?

There are many tools available to identify and mitigate bias:

- Data visualisation technologies can bridge the growing gap between digital data and human comprehension of that data. The Alan Turing Institute's Visualisation Interest Group is looking into how these can be created to visualise processes and bias in algorithms.
- During one of The Alan Turing Institute's recent Data Study Groups,[7] a group of talented early career researchers collaborated with Accenture to develop a very early version of a "Fairness Tool" to promote fairness in algorithmic decision-making across the financial services sector.
- IBM has developed *AI Fairness 360*, which is an open-source toolkit of metrics to check for unwanted bias in data sets and machine learning models, as well as algorithms to mitigate such bias.[8]
- *Fairness Measures* provides a tool to detect bias through several fairness metrics and access to datasets.[9]

---

[7] See https://www.turing.ac.uk/collaborate-turing/data-study-groups/accenture-challenge-fairness-algorithmic-decision-making

[8] Bellamy, R.K., Dey, K., Hind, M., Hoffman, S.C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A. and Nagar, S., 2018. Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. arXiv preprint arXiv:1810.01943. https://arxiv.org/pdf/1810.01943v1.pdf.

[9] See http://www.fairness-measures.org/

- *FairML* is an end-to-end toolbox for auditing predictive models by quantifying the relative significance of a model's inputs. It allows analysts to more easily audit complex predictive models that are difficult to interpret.[10]
- *FairTest* allows the user to test biases in a dataset by checking for associations between predicted labels and protected attributes. It also provides a way to identify regions of the input space where an algorithm might incur unusually high errors.[11]
- *Aequitas* detects bias by offering several fairness metrics, including demographic or statistical parity and disparate impact, along with a 'fairness tree' to help users identify the correct metric to use for their particular application.[12]
- *Themis* is an open source bias-identification toolbox that measures group discrimination and causal discrimination.[13]
- *Themis-ML* is a bias detection and mitigation tool that provides fairness metrics and some mitigation algorithms.[14]
- *Fairness Comparison* is an extensive library of bias-detection metrics and bias-mitigation methods that acts as a test-bed to allow different bias metrics and algorithms to be compared in a consistent way.[15]
- *FairLearn* by Microsoft is a Python programming package that implements a black-box approach to fair classification.[16]

---

[10] Adebayo, J. A. "FairML: Toolbox for diagnosing bias in predictive modeling". Master's thesis, Massachusetts Institute of Technology, 2016.

[11] Tramer, F., Atlidakis, V., Geambasu, R., Hsu, D., Hubaux, J.-P., Humbert, M., Juels, A., and Lin, H. FairTest: Discovering unwarranted associations in data-driven applications. In IEEE European Symposium on Security and Privacy, pp. 401–416, 2017. https://doi.org/10.1109/EuroSP.2017.29.

[12] Stevens, A., Anisfeld, A., Kuester, B., London, J., Saleiro, P., and Ghani, R. Aequitas: Bias and fairness audit, 2018, Center for Data Science and Public Policy, The University of Chicago.

[13] Galhotra, S., Brun, Y., and Meliou, A. Fairness Testing: Testing software for discrimination. In Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering, ESEC/FSE 2017, pp. 498–510, 2017.

[14] Bantilan, N. Themis-ml: A fairness-aware machine learning interface for end-to-end discrimination discovery and mitigation. Journal of Technology in Human Services, 36(1):15–30, 2018.

[15] Friedler, S. A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E. P., and Roth, D. A comparative study of fairness-enhancing interventions in machine learning. http://arxiv.org/abs/1802.04422.

[16] Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J. and Wallach, H., 2018. A reductions approach to fair classification. arXiv preprint arXiv:1803.02453. https://arxiv.org/abs/1803.02453.

- *Fairness in Classification* is a logistic regression implementation in the Python programming language for fair classification mechanisms.[17,18,19]
- *Procedurally Fair Learning* focuses on the fairness of the decision-making process and what features individuals consider to be fair during that process, rather than on the fairness of the outcome of a decision.[20,21]
- *Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness* provides a tool for identifying when an algorithm appears to be fair on an individual group, but is unfair on subgroups, and providing tools to mitigate this.[22]
- *Remove problematic gender bias from word embeddings* presents a way to remove gender stereotypes from word embeddings that are often used in machine learning and natural language processing tasks.[23]

---

[17] Fairness Constraints: Mechanisms for Fair Classification, Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, Krishna P. Gummadi. 20th International Conference on Artificial Intelligence and Statistics (AISTATS), Fort Lauderdale, FL, April 2017.

[18] Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment, Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, Krishna P. Gummadi. 26th International World Wide Web Conference (WWW), Perth, Australia, April 2017.

[19] From Parity to Preference-based Notions of Fairness in Classification. Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, Krishna P. Gummadi, Adrian Weller. 31st Conference on Neural Information Processing Systems (NeurIPS), Long Beach, CA, December 2017.

[20] Grgic-Hlaca, N., Zafar, M.B., Gummadi, K.P., & Weller, A. (2018). Beyond Distributive Fairness in Algorithmic Decision Making: Feature Selection for Procedurally Fair Learning. AAAI.

[21] Nina Grgić-Hlača, Elissa M. Redmiles, Krishna P. Gummadi, and Adrian Weller. 2018. Human Perceptions of Fairness in Algorithmic Decision Making: A Case Study of Criminal Risk Prediction. In WWW 2018: The 2018 Web Conference, April 23–27, 2018, Lyon, France. ACM, New York, NY, USA 11 Pages. https://doi.org/10.1145/3178876.3186138.

[22] Kearns, M., Neel, S., Roth, A. and Wu, Z.S., 2017. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. arXiv preprint arXiv:1711.05144. https://arxiv.org/pdf/1711.05144v5.pdf

[23] Bolukbasi, T., Chang, K.W., Zou, J.Y., Saligrama, V. and Kalai, A.T., 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, pp. 4349-4357.

- *Debiasing Representations by Removing Unwanted Variation Due to Protected Attributes* is an approach that researchers have applied to ProPublica's COMPAS dataset, which looks at the risk of offender recidivism in the criminal justice system.[24]
- The Algorithmic Justice League is working on an algorithm to challenge biases.[25]

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?

Credit scoring is a well-known tool which is used for more than just lending. Credit scores are consulted in many countries when renting, for example, and in some countries they are used in hiring decisions.[26] Currently, the core variables used in credit scoring are widely known, and many online services allow users to adapt their behaviours to improve their access to credit. As more diverse data sources are used, and stronger insights are gained regarding what constitutes a desirable or risky behaviour, it can be of societal interest to encourage communication of these results to users. These insights would serve as powerful nudges toward safer financial behaviours and would effectively serve as online credit counsellors. More research is necessary in this area, but evidence suggests that being better informed leads to better financial conduct.[27]

-End-

[24] Bower, A., Niss, L., Sun, Y. and Vargo, A., 2018. Debiasing representations by removing unwanted variation due to protected attributes. arXiv preprint arXiv:1807.00461, https://arxiv.org/abs/1807.00461.

[25] See https://www.ajlunited.org/

[26] Clifford, R., & Shoag, D. (2016). 'No More Credit Score': Employer Credit Check Bans and Signal Substitution.

[27] Collins, J. M., & O'rourke, C. M. (2010). Financial education and counseling – still holding promise. *Journal of Consumer Affairs*, *44*(3), 483-498, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1529422

Appendix

The following researchers contributed to this response:

Michael Rovatsos (Turing Fellow and Turing University Lead, University of Edinburgh)

Vaishak Belle (Turing Fellow, University of Edinburgh)

Ricardo Silva (Turing Fellow, UCL)

Cristián Bravo (Turing Fellow, University of Southampton)

Nick Holliman (Turing Fellow, Newcastle University)

Adrian Weller (Turing Programme Director for AI)

Aida Mehonic (Turing Programme Manager for AI)

Didem Özkul (UCL)

Alan Dix (Swansea University)

turing.ac.uk
@turinginst