**Fairness and transparency in the age of the algorithm**

*Never mind the power of algorithms, what about their probity?* ***Sofia Olhede*** *and* ***Russell Rodrigues*** *discuss recent efforts to ensure greater scrutiny of machine-generated decisions*

The algorithm might be considered the workhorse of the modern digital economy. No matter what business you are interacting with – whether in media, retail, healthcare or finance – there's a good chance that, behind the scenes, an algorithm is churning through your data, running through a series of steps or calculations, to automate or support a decision that the company needs to make. It might be a decision about what advert you see, or what product to recommend, or whether you qualify for a much-needed bank loan.

Reports from the UK and US governments point to efficiency savings as one of the potential benefits of using algorithms. [1,2] But there are potential costs too. Recent news stories have highlighted the risks of skewed outputs, and how they can entrench social inequalities.[3] Indeed, the December 2016 issue of *Significance* featured an interview with Cathy O'Neil, whose book *Weapons of Math Destruction* discusses this very issue.[4]

Given the broadening range of domains in which algorithms are applied, it is crucial to ensure that machine-generated decisions are fair and unbiased, especially when they affect human lives. The problem is that much of what an algorithm does is hidden from view – inscrutable to those whose data it feeds on.

How do we ensure fairness in the age of the algorithm?

**A peek inside the 'black box'**

While great technological strides have been made to enhance the capability of algorithms, efforts to define the ethical frameworks for implementation are only just beginning, though they are rapidly gathering pace.

'Transparency' is a concept mentioned frequently in these debates, and it is an important one: we cannot evaluate the probity, or fairness, of a computer-generated decision without first having a clear understanding of the population from which the

---

[1] Artificial intelligence: opportunities and implications for the future of decision making, GO-SCIENCE

[2] Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights, US Government, Executive Office of the President, US Government

[3] https://www.theguardian.com/technology/2016/dec/19/discrimination-by-algorithm-scientists-devise-test-to-detect-ai-bias

[4] B. Tarran (2016). Significance 13, vol. 6, pp 42-43.

data is drawn and the logical steps an algorithm goes through to determine its outputs. [5,6]

Even if we can identify the variables fed into an algorithm, data which reflect poor sampling design, unconscious bias, or which contain irrelevant correlations will have repercussions for the computed output: the algorithm can only work with the data supplied to it. Initiatives to improve the quality and availability of data, such as those championed by the UK Open Data Institute, can help to alleviate some of these issues. But better data will not solve all problems because – in many cases – an algorithm's inner workings resemble a 'black box', and it is often unclear precisely how the input data is used to reach a decision.

Modern machine-learning (ML) algorithms are typically designed and trained to excel in predictive accuracy using massive volumes of data. The availability of extremely large data sets, along with modern computational power, makes this approach quite practical. However, with prediction as the endpoint, such algorithms tend to assimilate the input data and construct complex models with convoluted and interacting components. This is especially evident with the intricate, multi-layered ML systems used in deep learning and convolutional neural networks. It thus becomes difficult to unpick specific strands of the decision-making process to understand precisely how a conclusion was reached.

By contrast, traditional statistical algorithms are concerned with explanation as well as prediction, and tend to use clearly-specified, often linear models, which are easier to scrutinise – although they are, on occasion, less powerful. In some instances, the impressive performance of ML algorithms can make the lack of transparency a reasonable trade-off, but this may not always be the case.

Without the ability to thoroughly scrutinise algorithms, there is little recourse when contested judgements are made. That is why transparency features so heavily in current discussion. Transparency may not seem important if an algorithm is simply recommending films and restaurants, but it is of greater concern where mortgage or employment applications are concerned. Few people are likely to be satisfied with a 'Computer Says No'-style rejection, should they fall foul of an algorithm into which they have no insight. Similarly, it is little comfort to be told: "We used mean square error as our error metric in prediction and unfortunately you do not qualify."

The challenge therefore is to make algorithms transparent, fair and intelligible to the people affected by their outputs.[2]

**Finding the right words**

---

[5] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. S. Zemel (2012). Fairness through awareness. In Innovations in Theoretical Computer Science, Cambridge, MA, USA, January 8-10, 2012, pages 214–226.

[6] J. Kleinberg, S. Mullainathan & M. Raghavan (2016). Inherent Trade-Offs in the Fair Determination of Risk Scores, arXiv1609.05807.

Early research efforts have focussed on developing a technical vocabulary to describe such problems.[5,6] Cynthia Dwork of Harvard University has proposed a framework which explicitly quantifies the notion of fairness with respect to a given population, thus ensuring it is built into the decision-making process. Further work in this direction will enable the balance between algorithm performance and the fairness of decisions to be measured objectively, and not simply described in abstract terms.

In addition, a number of scholarly societies are convening discussions and beginning to draw up ethical frameworks for algorithm design and implementation. The Association for Computing Machinery has issued a statement on 'Algorithmic Transparency and Accountability', with seven recognised principles to maximise scrutiny and minimise harm. Similarly, the IEEE Standards Association has launched a global initiative on the ethics of artificial intelligence and autonomous systems, highlighting the responsibility and accountability needed to ensure that algorithms do not infringe human rights. In the UK, the Royal Society and British Academy have established a working group comprising lawyers, philosophers, social scientists, mathematicians, statisticians and computer scientists. This group is considering the ways in which data-driven technologies can be best governed so as to reap the benefits of innovation whilst preserving integrity and trust in the eyes of the public.

**Open dialogue**

Societal attitudes are ultimately shaped through public discourse, and this will be especially true of attitudes towards algorithm-powered technologies. However, public understanding is currently hindered by a technical barrier. Deep-learning algorithms, for example, are a scientific frontier; even experts can seldom describe their mechanics in granular detail, and it is even more challenging to do so in a manner that allows fairness to be assessed and understood.

But public understanding is crucial: after all, a key principle of European data protection law is that of "informed consent" – that people give permission for their data to be used only once they understand exactly how and why it will be used. In the context of complex algorithms, it may be easy to explain what data will be used, but the technical detail of the processes involved may mean that the "how" and "why" are less easy to communicate.

This challenge will only become greater as algorithms become ever more sophisticated. Transparency is an important step to establish and maintain fairness, but a much broader framework of governance is required to ensure that algorithms are implemented responsibly, with proper accountability for and understanding of the decisions that are made.

**Author bios**

**Sofia Olhede** is a professor of statistics at University College London, and scientific director of the UCL Big Data Institute

**Russell Rodrigues** is operations manager of the UCL Big Data Institute

ENDS