AUTOMATED DECISION-MAKING IN THE PUBLIC SECTOR

This article analyses the use of automated decision-making (ADM) in the public sector, the potential problems it poses and the effects these may have in the context of public law.

by Rebecca Williams, Professor of Public Law and Criminal Law, and Thomas Melham, Professor of Computer Science, Oxford University and *Practical Law Public Sector*

RESOURCE INFORMATION

RESOURCE ID
w-028-6934

RESOURCE TYPE
Article

PUBLISHED ON DATE
8 December 2020

JURISDICTION
England
Northern Ireland
Scotland
United Kingdom
Wales

While discussion of the A-level results over the summer may have brought the idea of formulaic, algorithmic, or automated decision-making (ADM) by public authorities to the forefront of public consciousness, in fact there is a wide range of circumstances in which public authorities can and do deploy ADM.

An investigation by The Guardian in 2019 showed that some 140 of 408 local authorities in the UK were using privately developed algorithmic "risk assessment" tools, particularly to determine eligibility for benefits and to calculate entitlements (*One in three councils using algorithms to make welfare decisions, theguardian.com, 15 October 2019*). Although there has apparently been some decline in the popularity of these systems recently (*Councils scrapping use of algorithms in benefit and welfare decisions, theguardian.com, 24 August 2020*), their attraction is clear. As the Data Justice Lab's Project Report on *Data Scores as Governance: Investigating uses of citizen scoring in public services (December 2018)* points out, many such systems were developed "in part as a response to on-going austerity measures" (*at page 74*) and austerity formed a "recurring theme" in their investigations into the rationale for implementing such data systems (*at page 116*). If a system can replace or support a decision-making process more quickly, more efficiently, and much more cheaply than a human decision-maker might previously have done unaided by technology, it is not surprising that a public authority should choose to adopt it. Indeed, it is not just that such systems can supplant human decision-making; they also have the potential to improve on it, as has already happened in the context of medicine (*Mayer McKinney, Sieniek, Shetty et al: International evaluation of an AI system for breast cancer screening, Nature (577), 2020, at pages 89-94*). Of course, the opposite was true in the context of A-level results.

Although such systems have the capacity to bring huge benefits, they also carry several risks. In order to understand these risks, and how the law might respond to them, it is necessary first to take a step back and get a sense of how these systems operate.

AI: AN INTRODUCTION

Artificial intelligence (AI) is often said to be of two kinds, each with different aims. The first is "Artificial General Intelligence", sometimes called Strong AI, which can learn to do any intellectual task that human can do. This is the stuff of science fiction, and many scientists believe that it will be decades before general AI is attained, if ever. The other kind, "Applied AI" is more relevant here and has the more modest aim of producing systems that can succeed in one specific task that is usually thought to require human intelligence. AlphaGo, the computer program

that learned to be a champion at the board game Go, is one prominent example. It is this kind of AI, of course, that is most relevant for public lawyers' understanding of ADM.

There are many technical approaches to Applied AI, each with their own strengths and weaknesses, and characteristics such as the degree of interpretability. The Information Commissioner's Office (ICO) and the Alan Turing Institute have jointly produced guidance, *Explaining decisions made with AI (May 2020)*, which has a useful and relatively accessible summary of these (*at Annexe 2*). Methods can be broadly classified as taking one of two distinct approaches, though hybrid methods also exist:

- **Logical AI.** This represents the fundamental logic of decisions, to some extent explicitly, such as a decision tree or a set of explicit rules. This is relatively interpretable AI and has a long history.

- **Statistical AI.** This takes a very different technical approach and is, perhaps, the AI currently most prominent in the public mind. In Statistical AI, specialised algorithms are used to create complex statistical models of the factors that go into making a decision, based on data from a large number of already determined cases.

A prominent general method for building Statistical AI models is "supervised learning". The system learns a model by being presented with a large number of input-output pairs from a database of previously made decisions, such as decisions made, for example, by a human judge or administrator. This is called the "training data" and is used to train the model to correctly indicate the already-known decision outcomes. The inputs are the "features" of the case being decided (such as a prosecution or claimant's evidence) and the output is the decision that has been made (whether or not the claim was upheld, for example).

A supervised learning algorithm analyses the training data and produces an inferred mapping from inputs to outputs, which can then be used for "unseen" cases outside the training set. In the best case, the model will be able to correctly determine the most appropriate decision outcome for the vast majority of unseen cases. (In AI parlance, this is sometimes called making "predictions", though this does not have the normal connotation of saying that something is going to happen in the future.) For this to be successful, the learning algorithm must generalise beyond the training data to unseen cases in a reliable way. It is this capability of the trained model that is used to support or even automatically execute decision-making.

RISKS AND UNKNOWNS

However, care must be taken when we speak of a statistical model making "'correct" predictions. Correct according to what criterion? To assess this, there must be some source of authoritative "ground truth". Usually, some further data not involved in the training but also with known outcomes is used to assess the accuracy of prediction. But accuracy can also be measured in several different ways, and the choice has to fit the ethical and social requirements for the decision-making application being supported. The risks which such systems might give rise to include:

- As various authors have pointed out some (though not all) systems operate as "black boxes", meaning that the rationale for the decision made or suggested is opaque or invisible (see, for example, Cathy O'Neill, *Weapons of Math Destruction* (Penguin, 2017) and Viktor Mayer-Schönberger and Kenneth Cukier, *Big Data* (John Murray, 2013)). This poses an obvious threat to the public law requirement that in many circumstances reasons should be given for decisions and notice should be given of the case against an applicant (*R (Institute of Dental Surgery) v Higher Education Funding Council [1994] 1 WLR 242*; *Mark Elliot: Has the Common Law Duty to Give Reasons Come of Age Yet?, Public Law (56), 2011*).

- There is a real danger that the prospect of decreased costs of the decision-making process raises the level of tolerance for a decreased level of accuracy.

- As a *2018 Report* by the AI Now Institute at New York University points out, "while individual [human] assessors may also suffer from bias or flawed logic, the impact of their case-by-case decisions has nowhere near the magnitude or scale that a single flawed automated decision-making system can have across an entire

population" (*at page 18*). This might not be so problematic were it not for the fact that, as O'Neill's *Weapons of Math Destruction* demonstrates, such systems may actually create the problems they are intended to spot. For example, if a system predicts that an individual is likely to re-offend and that individual is consequently imprisoned rather than being given a community sentence, that can itself increase the chances of that person re-offending.

This latter point is, of course, not a problem exclusive to ADM, or even one which is exclusive to a digital form of ADM rather than a rigid policy applied by human decision-makers, but it is a further issue that ADM has the capacity to scale up. In addition, ADM by definition cannot make exceptions other than those it has already been trained to make.

As the A-level events demonstrated, the cumulative result of these different risks is a further exacerbation of existing information asymmetries. Existing injustices (such as, as was a factor in the context of A-level results, attendance at a poorly performing school) may be entrenched or worsened by the ADM system. It is not that these difficulties cannot occur in the context of purely human decision-making, but simply that while a human decision-maker has the capacity to spot additional features, today's automated systems will only ever examine the features they were initially configured to examine. The model may behave accurately on its own terms, but if it has not been configured to deal with every possible combination of events that may occur, its application to a particular case may be problematic. This may increase the potential for mistakes to occur, and certainly increases the scale of their impact when they do.

THE RISE OF ADM JUDICIAL REVIEWS?

It seems likely, therefore, that the use of such systems will become the subject of judicial review actions, and that public law in the 21st century will have to become as familiar with reviewing ADM as it is with reviewing human decision-making. Such actions were threatened both in response to the A-level results decisions, and in challenging a Home Office algorithm that filtered UK visa applications. While both those systems were abandoned before cases on them could be heard, a judicial review challenge to the police's use of automated facial recognition (AFR) was successful in the Court of Appeal (*R (Bridges) v Chief Constable of South Wales Police (Respondent) and others [2020] EWCA Civ 1058*). In Bridges, the court held that there was insufficient guidance on where the technology could be used and who could be put on a watchlist, that a data protection impact assessment was inadequate, and that a police force had not taken reasonable steps to investigate whether the technology had a racial or gender bias as required by the Public Sector Equality Duty (PSED) (under the Equality Act 2010). For further information, see *Legal update, Facial-recognition technology in breach of human rights and data protection legislation (Court of Appeal)*.

As such systems are deployed and, therefore, challenged more widely, it seems likely that they will give rise to a series of legal issues, including:

- Several decisions are made in the process of setting up such a system, each of which can in principle be subject to challenge. As the ICO and Alan Turing Institute's *joint guidance* states, there is a wide variety of ADM systems that might be used and public authorities may well be challenged on their decision to use one such system rather than another.

- The system may well need to be trained and choices will have to be made about the data used for training.

- The system will need to be tested, and again choices will need to be made between these different measurements.

As is clear from *Bridges*, there is, at least under the PSED, a burden on public authorities to ensure that they are informed about how precisely their ADM systems have been trained and tested. If it is felt that the wrong measurement has been used, then their use of a particular system will be unlawful. On this basis it seems at least probable that in future cases the court may well go beyond the requirements of the PSED and also make use of the existing grounds of judicial review (I will shortly publish an article, *Algorithmic Decision Making and Judicial Review*,

which relates to this). In *Bridges* it was held that even the presence of a "human in the loop" was not sufficient to fulfil the PSED, but it seems likely that similar discussions will need to take place in relation to the question of whether or not a decision-maker has unlawfully delegated its power or fettered its discretion (see *Practice note, Judicial review: an introduction: Grounds for judicial review*).

The ICO and Alan Turing Institute's *joint guidance* is focused on ensuring the "explainability" of such systems. It seems likely that this too is an area that will see challenges, either to the choice of one particular less explainable form of ADM over another, or regarding the lack of explanation given in a particular instance, on grounds such as the duty to give reasons or notice of the case against an affected individual (see *Practice note, Duty to give reasons*). In addition, the particular factors taken into account by a system, its overall reasonableness (in the *Wednesbury* sense), or even its proportionality, may well give rise to challenges too. For each of these grounds, as in *Bridges*, it will be vital for the court to have a clear understanding of how the system works, how it has been trained and what measurements have been used to assess its operation, so that public law can start to develop a framework to deal with these issues in the way that it has for situations involving purely human decision-making in the past.

SAFEGUARDING UNDER THE GDPR

Some assistance on this front comes from the General Data Protection Regulation (*(EU) 2016/279*) (GDPR), which under Article 22 states that data subjects have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning them or similarly significantly affects them unless it has a legal basis (see *Practice note, Overview of GDPR: UK perspective: Automated decision making (including profiling)*). Although a data subject's consent can in principle provide such a legal basis, given recital 43, it seems unlikely that this will work for public authorities given the "clear imbalance of power" between them and individuals and private entities (see *Practice note, GDPR and DPA 2018: implications for the public sector: Consent*). This suggests that automated processing must instead be authorised by member state or EU law that also lays down suitable measures to safeguard the data subject's rights, freedoms and legitimate interests. Where such processing deals with special category data (under Article 9) processing must be necessary for reasons of substantial public interest and must be proportionate and provide for suitable and specific measures to safeguard the fundamental rights and interests of the data subject (*Article 22(4)*). For further information, see *Practice note, Overview of GDPR: UK perspective: Special categories of personal data.*

Articles 13(2)(f), 14(2)(g) and 15(1)(h) of the GDPR also give data subjects the right to know of the existence of ADM, including profiling, and "meaningful information about the logic involved", as well as the significance and envisaged consequences of such processing for the data subject. Precisely what "meaningful information about the logic involved" might entail has already been the subject of a lively academic discussion (see *B Goodman and S Flaxman: EU regulations on algorithmic decision-making and a right to explanation, AI Magazine (38), 2017; S Wachter, B Mittelstadt and L Floridi: Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation, International Data Privacy Law, 2017; A Selbst, J Powles: Meaningful information and the right to explanation, International Data Privacy Law 233, 2017*). The European Data Protection Board (formerly the Article 29 Working Group) has suggested that the information provided should be "sufficiently comprehensive for the data subject to understand the reasons for the decision" such as "details of the main characteristics considered in reaching the decision, the source of this information and the relevance" (*Article 29 Data Protection Working Party: Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679, at page 26*). For further information, see *Practice note, GDPR and DPA 2018: profiling and automated decision-making (UK)*.

In terms of the domestic application of this legislation, the ICO and Alan Turing Institute's *joint guidance* has assumed that from January 2021 and the end of the UK-EU transition period, references to the GDPR should be read as references to the equivalent articles in the UK GDPR (at page 10), and the government has published a *Keeling Schedule to the GDPR*, though it now remains to be seen how precisely the legislation will work alongside the government's recently published *national data strategy*. But in any event similar provisions are contained in section 14 of the Data Protection Act 2018, which states that where a significant decision, based solely on automated processing, has been required or authorised by law, the data controller must as soon as reasonably practicable notify the data subject of this. The data subject then has one month to request the controller to

reconsider or take a new decision which is not based solely on automated processing. In addition, the Secretary of State may by regulations make further provisions to safeguard the data subject's rights, freedoms and legitimate interests in this context.

It therefore seems likely that in future courts will not only have to grapple with the implications of such systems for the standard public law grounds, but also to interpret other pieces of legislation in the way that they have already done in relation to the PSED. In order to do this, it is vital that computer scientists and lawyers work together as effectively as possible so that courts can develop a better understanding of how these systems work and thus how public law can best be deployed to provide an optimal form of control.