Legal and regulatory frameworks governing the use of automated decision making and assisted decision making by public sector bodies

Workshop briefing paper

Lilian Edwards
Rebecca Williams
Reuben Binns

July 2021

The Legal Education Foundation

Contents

Introduction

This technical legal briefing paper aims to:

- Explain the existing legal and regulatory frameworks governing the use of Automated Decision Making and Assisted Decision Making by public sector bodies;
- Identify gaps in the existing frameworks; and,
- Highlight any issues/problems that have been created by these gaps.

Section 1 outlines the different types of automated decision-making / decision-assisting systems and provides examples of their uses in the public sector. In Appendix A, some of the major technical issues, limitations and risks of these systems are also laid out.

Sections 2 and 3 presents the current legal framework, beginning with data protection law

Section 4 continues to analyse the current legal framework, considering equality law, and public law.

Section 5 covers some alternative regulatory models that feature in recently proposed AI / algorithm regulation, to put on the table some regulatory possibilities for fixing the gaps in the current frameworks.

1. Automated Decision-Making / Decision-assisting Systems in the Public Sector

There are a wide variety of automated decision-making (ADM) and assisted decision technologies (ASDM) in use by public bodies today. Exactly what counts as an ADM system is subject to debate, but we can draw a distinction between:

1. technologies which digitise or automate manual processes (e.g. providing an online digital form instead of a paper equivalent, or automatically disbursing a scheduled payment); and
2. those which actually affect a decision outcome in some way e.g Universal Credit.

The latter might involve a rule-based system or a statistical model.

Rule-based systems may be based on the formalisation in code of policies or legal criteria. The use of 'bright line' criteria for decision-making may therefore appear amenable to automation through this kind of system.

Statistical models are based on empirical patterns in historic data, and are used for classification and prediction; as such, their outputs are, by nature, not 100% accurate. Unlike rule-based systems, where the output might be logically implied by the inputs and the rules, statistical systems aim to provide the most likely answer, given previous cases. They are therefore typically deployed in situations where there is some inherent uncertainty, where some of the relevant facts of the case are unobservable or have not yet happened. Examples therefore include risk scoring and fraud detection.

In many cases, the two types of systems might be incorporated into a single solution; for example, with a statistical model being used to assess the risk of fraud, and that risk score being used as an input to a rule-based system for determining whether to trigger an investigation. It is also worth noting that recent systems for speech transcription, translation, or optical character recognition are also based on statistical models. While these are also not 100% accurate, their purpose is not to predict or classify cases as such, but rather to automate some previously human task using statistical methods. Various formerly manual processes are becoming partially automated through such systems.

Statistical models have existed in various forms for centuries, but recent computational methods and availability of data have made it far easier to construct more complex statistical models with much larger sets of data. The kind of statistical models previously confined to sectors like insurance, based on large, high-dimensional datasets, are now feasible to deploy in a wider array of organisational contexts, including in the public sector.

Both rule-based and statistical models can be used as decision-support tools, to assist human decision-making, or to make decisions automatically. The boundary between automating and assisting decision-making can be tricky to define and uphold in practice, (as discussed in section 3.4 below).

Examples of the use of algorithmic decision-making systems in the public sector abound. They include:

- Risk scoring individuals for prioritising interventions in social care[1]
- Assessing risk in child welfare contexts[2]
- Predicting the outcomes of potential inspections based on historic data, in order to better target resources for in-depth inspections. This approach is

[1] Dencik, Lina, et al. "Data Scores as Governance: Investigating uses of citizen scoring in public services." *Cardiff: Data Justice Lab* (2018).
[2] Ibid

taken by Ofsted in relation to school inspections[3] and the Department for Transport / DVSA in relation to MOT test centres.[4]
- Using natural language processing (based on machine learning) for classifying documents into categories, or classifying feedback from users of government services.[5]

Such prominent, publicised cases are examples of ADM systems built **in-house** by digital teams working within central government departments or local authorities, often using training data drawn from within existing government services. However, ADM is also typically integrated through various degrees of **outsourcing**, including to private sector organisations and reliance on paid and open source software tooling.

These different models for ADM development and integration have different trade offs. The in-house approach can be more costly up front with a large investment of staff requirement required but allows more control and oversight over the whole process. Outsourcing can be (potentially) cheaper upfront but purchasers may have less control to make changes; can become locked into proprietary models which may link to issues of intellectual property.

In addition to these directly commissioned and explicit uses of ADM, these technologies find their way into the public sector in subtler, less direct ways. Common integrated software-as-a-service enterprise IT services, such as Microsoft's Office 365, are increasingly incorporating 'AI' into their standard suite of software. For instance, a familiar example would be spam filters. These have been offered as standard in email services; these are based on statistical models designed and built by the provider, often trained on data drawn from clients. While largely ignored in debates about the use of algorithmic decision-making in the public sector, they are a useful example of how such systems can become completely embedded and accepted to the point of being almost invisible. However a spam filter which marks legitimate email queries as spam based on behavioural or geographical features of the content or sender could result in the sender's message being excluded from important services and opportunities.

Regardless of what kind of ADM systems are procured and how they are integrated into the public sector, there are a variety of now well-known limitations and potential harms that can arise from their use in decision-making. **Appendix A provides an overview of some of the issues and risks which should be discussed and mitigated for when developing and deploying ADM systems**. These include:

- Error
- Discrimination and equality issues
- Robustness, generalisation and feedback loops
- Limits to prediction

---

[3] Risk assessment methodology: good and outstanding maintained schools and academies
[4] How the Department for Transport used AI to improve MOT testing
[5] Natural Language Processing in government - Data in government

- Individual level accuracy
- Unobserved labels
- Opacity, transparency, explainability
- Correlation vs causation
- Automation bias, rigidity and over delegation

The remainder of this report focuses on explaining the existing legal and regulatory frameworks governing the use of Automated Decision Making (ADM) and Assisted Decision Making by public sector bodies.

2. Current legal framework: Overview

The key laws involved in regulating public sector ADM systems currently come from

(a) Data protection (DP) law
(b) Equality law including the Public Sector Equality Duty and
(c) The common law of judicial review

A number of other laws may be involved in governance of ADMs, including, notably, labour law. For example, there is concern that automated hiring systems (and hiring triage, firing, promotion, disciplinary management systems etc) may be discriminatory and infringe labour standards[6]. However these issues arise equally in the private and public sectors and in this briefing, we do not feel these issues are particularly germane to *public sector* ADM.

Contract law, between public body and system developers, or between service provider and service user, is also often relevant. In particular, the rights of public bodies commissioning ADM systems from private vendors may often be dependent on contract and it will usually be formulated via procurement frameworks. We raise some issues around these in section 4.4.

In terms of the current law applicable to public sector ADM systems, it is perhaps helpful to think of the legal issues divided into a series of stages;

1. Can/should an ADM system be undertaken at all? Should some types of automated decision making in particular domains simply be unlawful?
2. If ADM is lawful, what are the constraints on its development and deployment in a particular circumstance?
3. What rights do users have to challenge decisions reached using an ADM system, and when, and how?
4. What remedies might be available after such a challenge?

These questions are not neatly answered by one area of law, and sometimes by none at all. There is a complex and context-dependent matrix of rules. We try at

[6] See Denzik, Sanchez and Edwards “What does it mean to solve the problem of discrimination in hiring? Social, technical and legal perspectives from the UK on automated hiring systems”, Conference on Fairness, Accountability, and Transparency (FAT* '20), January 27-30, 2020, Barcelona, Spain: arXiv:1910.06144 [cs.CY].

the end of this section to present in table form how key problems are addressed by different legal instruments, and how well.

3. Current legal framework: GDPR

3.1 Key issues

The GDPR is probably the law most commonly discussed in relation to regulating public (and indeed private sector) ADM. Since many or indeed almost all public sector ADM systems process personal data[7], one of the most obvious legal regimes invoked to regulate ADMs is data protection (DP)[8]. However, many problems arise when it is applied to this domain, which make it wholly insufficient alone for this task. **Issues include:**

Responsibility for DP obligations : controllers, processors, joint controllers

- **It is not clear what role as *controllers, joint controllers or processors,* multiple private sector vendors and/or multiple public bodies may play**, in relation to provision of training set data, models, algorithms, or sub-decisions. Yet this classification is crucial to understanding what GDPR obligations are owed to users, by whom and when.

Use of Sensitive Personal Data (SPD)

- **It is not always clear when SPD can lawfully be processed or created by public sector ADMs** (including SPD "revealed" or inferred in the context of machine learning systems), and what safeguards should be put in place if it is.

User Rights and scope of Art 22

- Although the GDPR does give users active rights to, for example, erase their data, or object to profiling using their data, **these rights are largely**

---

[7] Personal data is any information relating to an identified or identifiable individual (GDPR, art 4(1)). It can include names, addresses (both physical and Internet Protocol addresses), ID card numbers, as well as email addresses and other data relating to individuals. The GDPR refers explicitly to "a name, an identification number, location data, an online identifier, or by reference to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person". It is important to note that pseudonymised data remains personal data (GDPR art 4(5)) and that data held by more than one data controller which potentially make a data subject identifiable with reasonable likelihood, subject to constraints such as the time and money needed for identification, fall within the remit of personal data (recital 26).

[8] The useful Australian report Montoya D and Rummery A "The use of artificial intelligence by government: parliamentary and legal issues" NSW Parliamentary Research Service, September 2020, e-brief 02/2020 at https://apo.org.au/sites/default/files/resource-files/2020-09/apo-nid308503.pdf reports that in its survey of public sector AI regulation, the GDPR was the most commonly cited example.

useless in the public sector context where users have to engage to get the services they want, and the state is a monopoly provider.

- Because of the above, much emphasis has been placed on the right in **art 22 of the GDPR** to object to decisions made solely by automated processing ("solely automated decisions" or SADs), with a view to obtaining safeguards such as having the decision taken by a human. **However it is very unclear what the *scope of GDPR art 22* is**.
- European Court of Justice (CJEU) and UK **case law on art 22 are to date non-existent so uncertainty persists** and even recent national EU case law providing guidance on ADMs in art 22 as well as the information rights in GDPR arts 13-15 is partial and unhelpful. **Development of case law in the UK would help in reducing uncertainty in this area, as ICO and EDPB guidance, though useful, remains simply that. Failing that however, more detailed statutory definitions of these terms in the post Brexit era, might help, as might sector-specific codes.**
- **Particular uncertainty relates to the notions of a "decision", "solely automated" and "significant effects".** It is likely that many if not most public ADM systems may be easily excluded from the scope of art 22, whether deliberately or accidentally, thus minimising its perceived protections. **These uncertainties about scope may mean that art 22 without substantial amendment or reinterpretation by case law is "beyond saving" in relation to ADM.**
- **Even where art 22 is in scope, it is unclear if art 22 gives users a *"right to an explanation"* of how the system affects them.**

Transparency Rights- Arts 13-15

- **Arts 13-15 of the GDPR** may provide alternative routes to obtaining "**meaningful information about the logic involved**" in an ADM. However even then problems arise such as
  - (i) How to obtain and convey this meaningful information in practical, technical and contractual terms
  - (ii) Possible restriction of the rights in arts 13(20(f), 14(2)(g), 15(1)(h) to art 22 SADs, thus reintroducing all the scope problems canvassed above
  - (iii) Restriction of rights by reference to the rights of third parties
  - (iv) Restriction of rights to information by trade secrets and other IP rights where systems have been bought wholly or partly from private vendors

Bias, discrimination and fairness in the GDPR

- **DP law in general is not best fitted to address *substantive issues of bias and discrimination* in public sector ADM** but must be looked at in this area as a minor partner to equality, human rights and public law (see section 4)
- **While *users* have the right to be informed of the existence of automated decision making, there is *no general requirement in the law to publish the existence of public sector ADMs*** and the lack of

such a register is a first hurdle to regulator or civil society oversight of such systems.

- **Data Protection Impact Assessments (DPIAs) are however highly useful in this area**. However, although a DPIA is highly likely to be required for a new public sector ADM system or adjustment of existing system, ***requirements of publication and meaningful user consultation*** **still do not exist in law**, though are often expressed in guidelines. These and other issues need to be addressed and we point in section 5 to **the proposed EU AI Regulation** as containing a number of potential, if heavily aspirational, solutions.

Processing of personal data within the UK currently remains governed by the General Data Protection Regulation (GDPR), even after Brexit, and as further implemented into UK law by the Data Protection Act 2018 (DPA 2018)[9]. The GDPR as a Regulation initially operated in the UK without need for transposition, but post Brexit continues to operate as the "UK GDPR"[10]. Notably, the Information Commissioner's Office (ICO) continues to regulate the regime and GDPR recitals continue to have the same status as before – they are not legally binding, but they do clarify the meaning and intention of the articles.

3.2 Responsibility for DP obligations: controllers, processors, joint controllers

The key obligations in the DP regime are found in the art 5 principles. Personal data must be processed lawfully, fairly, transparently, and for specific, explicit, legitimate purposes identified by data controllers. Processing is lawful if, and only if, at least one of the grounds listed in art 6, GDPR applies (consent, necessary for contract, compliance with a legal obligation, vital interests of a person, public task, legitimate interests of controller). For the public sector, art 6 (e), "processing is necessary for the performance of a task carried out in the public interest or in the exercise of official authority vested in the controller" ("public task") will usually be appropriate, and legitimate interests is excluded as a ground as, in all probability, is consent. The reason for this is that if a data controller is relying on consent, they must ensure that such consent is freely given, and fully informed. As such, consent cannot be relied upon "where there is a clear imbalance between the data subject and the controller." (recital 43) The classic example of this is where the data controller is a public authority or an

[9] In relation to law enforcement activities, the relevant parent directive is the DP Law Enforcement Directive 2016/680 which applies to "competent authorities" (much wider than the police – see s 30 of the 2018 Act but note excludes the intelligence services) and is transposed into UK law by the DPA 2018, Part III. Part IV of the DPA 2018 also applies similar but less extensive laws to the work of the intelligence services (which is a purely UK innovation as they are excluded from EU competence). In general, in this part of the briefing we assume the focus is the public sector excluding the criminal justice, law enforcement, and intelligence services sectors.

[10] ICO Information rights after the end of the transition period – Frequently asked questions

employer processing the personal data of employees. Thus ICO guidance states that consent would be an inappropriate ground to rely on for a public authority[11].

Obligations under the GDPR are in principle allocated in relation to an ADM system depending on who is the data controller or data processor, with historically the bulk of obligations falling on the former[12]. In the complex outsourcing situations often typical of public sector ADM (see Appendix 1, eg, buying in cloud services, training data, trained or partly-trained models, etc) there are increasingly knotty issues involved in discerning who is a data controller, a mere data processor or a *joint* data controller, for what stages of data processing and what the corresponding obligations are that flow from these categorisations. In an influential chain of CJEU case law, it has become apparent that a body may be deemed controller even though it does not have actual access to data processed, so long as it eg organises or coordinates the collection of that data (*Jehovan todistajat* C-25/17). Conversely a cloud service provider acting for a public body as a data processor might find itself deemed a joint controller, at least for certain stages of processing (GDPR art 26 and *Fashion ID* Case C-40/17). This has implications for who is responsible for GDPR obligations, particularly of information and transparency as well as responding to user rights such as rectification, erasure, and to object to profiling and solely automated decision making (see below**. Thus a public body must consider very carefully what relationship of controller/processor it puts into a contract with a supplier or service provider and what other rights it needs eg to access data, training sets, source code so that it may be able to meet its obligations under the GDPR. These matters need urgently to be considered in procurement frameworks (section 4.4).**

3.3 Sensitive personal data (SPD)

Special category data under the GDPR, in the UK, typically called “sensitive personal data” or SPD, is personal data so intimate it deserves extra protections (GDPR art 9). There is an exhaustive list of what constitutes SPD which includes revealing racial or ethnic origin, political opinions, or religious beliefs, or concerning an individual’s health, sex life or sexual orientation. It also includes biometric data but only where that is used to uniquely identify a person rather than, eg, categorise them as part of a group, eg, persons of colour, or mask-wearers. As is obvious, many key public sector systems will process SPD.

In order to process SPD, in addition to having an art 6 lawful ground for processing, a data controller must meet one of the conditions for processing outlined in Article 9(2) GDPR. If a data controller is relying on consent in this situation, it must be “explicit consent” meaning it must be explicit as well as freely given, specific, affirmative and unambiguous (GDPR art 4(11) and 7). Once

[11] ICO “When is consent appropriate?”

[12] A data controller is the individual, or body, who determines the means and purposes regarding the processing of personal data (Article 4(7) GDPR). Where a data controller determines the purposes and manner in which personal data are to be processed, a data processor is any body who processes that data (Article 4(8) GDPR).

again, this consent cannot be given in situations where there is a significant power imbalance, suggesting that public bodies should not be relying on it. Instead they should rely on one of a number of relevant grounds for the public sector, notably "substantial public interest" (art 9(2)(g) which is defined in DPA 2018, paragraphs 6 to 28 of Schedule 1, to include 23 "substantial public interest conditions". Other grounds which may be appropriate and relevant to public authorities include art 9(2)(h) (health or social care) and 9(2) (i) (public health). In each of these cases, the authority would have to show the existence of a legal power in UK law for their processing, and usually, produce an "appropriate policy document".

These detailed restrictions around SPD (particularly in relation to solely automated decision making, below) are a very live issue for machine learning systems because **arguably almost any such system may take ordinary personal data as inputs but "reveal" SPD by virtue of algorithmic induction or prediction** (for example, a health system that used diet or occupation data to predict health conditions, or a welfare system that used postcode and salary to partially output risk predictions concerning children in a household). The categorisation of a potential inference as SPD depends, according to the ICO, on how certain that inference is, and whether the system deliberately intended to draw an inference relating to one of the special categories[13].

In addition to cases where the inference *itself* may be SPD, there may also be cases in which a machine learning model indirectly infers SPD via a proxy as an intermediate step to performing some other kind of non-SPD inference. For instance, a CV filtering tool which has been trained on historic data reflecting sexist hiring decisions might indirectly infer gender from an applicant's education institution or word choices, and use that to make a prediction about their likely success if hired. **This would still constitute processing of SPD, even if it is inadvertent; as such, the ICO recommends proactively assessing this possibility and ensuring a lawful basis is in place to cover it if so**.

3.4 User rights and Article 22 GDPR

Users of systems which process personal data have a number of important rights under the DP regime, including rights to information (art 13 and 14), subject access rights (art 15), rights to erasure and restricting processing (art 17-19) and data portability (art 20). Under art 21, a user of an ADM where processing is justified under public task (art 6(e)) can object expansively to such processing, including profiling, on "grounds relating to his or her particular situation". **However such objection is of little gain if engagement is necessary eg to gain a welfare benefit or access health services.** In practice most attention in relation to GDPR user rights as tools to control ADM has focused on art 22 and to a lesser extent arts 13-15 (section 3.5).

[13] ICO Special category data

Art 22 of the GDPR 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. Each of the elements here is subject to much contestation. In the context of public sector ADMs, art 22 is a key focus of interest for at least four reasons[14]:

(a) Is it lawful for a decision to be taken in a solely automated way at all?

(b) Does a user have a right to object to such a decision and demand it be made by a human instead ("human in the loop")?

(c) If the user does not have such a right, what other safeguards must operate?

(d) In relation to (c) is one of these safeguards that the user has a "right to an explanation" of how an automated decision was made? If such a right cannot be derived from art 22, can it be found elsewhere in DP law? The question of explanations is significant in light of debates highlighted in Appendix A about (i) the "black box" nature of ML systems and (ii) the potential for error, bias and discrimination.

3.4.1 Scope of art 22

Art 22 has a very limited scope and these limitations have been barely explored in EU and UK jurisprudence. There is no extant CJEU case law, and very little member state case law. Even in relation to art 22's predecessor, art 15 of the Data Protection Directive, there is very little domestic case law. Only one state's courts, in the Netherlands, have so far issued decisions on art 22 rights since the GDPR and they relate not to public sector ADM but to the private sector and Uber and Ola ride-share drivers[15]. Many organisations have drawn primarily on the guidance of their DP regulator and the Art 29 Working Party (A29 WP), now the European Data Protection Board (EDPB), who addressed ADMs post GDPR in 2018[16].

---

[14] See further Edwards L and Veale M "Slave to the Algorithm? Why a 'Right to an Explanation' Is Probably Not the Remedy You Are Looking For" (2017) 16 Duke Law & Technology Review 18 .

[15] The "Uber" and "Ola" decisions, 11 March 2021, Amsterdam District Court; see unofficial English translations at https://ekker.legal/2021/03/13/dutch-Court-rules-on-data-transparency-for-uber-and-ola-drivers/ . See early discussion at Gellert R et al "The Ola & Uber judgments: for the first time a court recognises a GDPR right to an explanation for algorithmic decision-making" . Another art 22 decision was granted against Uber by default in somewhat unclear circumstances (Rb. Amsterdam - C/13/696010 / HA ZA 21-81) ; see https://gdprhub.eu/index.php?title=Rb._Amsterdam_-_C/13/696010_/_HA_ZA_21-81 .

[16] A29 WP Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679, revised and adopted 6 February 2018, 2018 17/EN WP251rev.01. See discussion on the original draft in Veale M and Edwards L "Clarity, surprises, and further questions in the Article 29 Working Party draft guidance on automated decision-making and profiling" (2018) 34 Computer Law & Security Review 398-404.

3.4.1.1 “Solely”

First, art 22 applies only when the decision has been based “*solely*” on automated processing (a SAD or solely automated decision). Almost invariably however, public sector ML systems that affect people’s lives significantly are not fully automated. One good analogous example from the law enforcement sector is the automated facial recognition system in *Bridges v SW Police* [2020] EWCA Civ 1058. Although the AFR Locate system utilised there was fully automated in making positive or negative matches of captured CCTV images to a list of images of suspects, the court accepted that no decision was ever made *deriving* from such a positive match without the intervention of at least one human (para 184); a “human fail safe”. However there was little enquiry in that case into whether these humans actually understood how the system operated or had access to its underlying training set data or algorithm, and thus it is questionable whether they had the ability to question or critique its findings (see discussions below re proprietary systems, trade secrets and meaningful human input).

It is obviously easy to introduce a nominal human into the loop, “rubber stamping” automated decisions, with a literal interpretation effect of knocking out art 22 rights. In reality though, there is some evidence that even where systems are explicitly intended only to support a human decision maker, for reasons of trust in automated logic (“automation bias”), lack of time, convenience or whatever, the system then tends to *de facto* operate as solely automated. These worries are all the more pertinent for public sector systems given their frequent lack of resources following austerity and the pandemic, especially as automated systems are often introduced specifically to allow reduction in skilled human time.

The A29 WP opined in 2018 that “meaningful human input” is required, rather than a “token gesture” for the system to be categorised as not “solely” automated”. The human overseer must be in the position to independently evaluate the case, assess the outputs of the system, and not simply rubber stamp them; they must have the authority to overturn the automated outputs; and they should be able to consider additional information and mitigating factors. This approach was accepted by the French DPA the CNIL in 2017 in respect to university admission[17]. In systems which are bought in from, trained by, or operated by, third party vendors, especially where downstream systems have contributed inputs to upstream decisions, this may be particularly hard to either prove or disprove. Even in the *Bridges* case, though it was not discussing art 22, the Court noted that the officer acting as a human failsafe was not a technical expert in the relevant software (or indeed in data science), casting doubt on the extent to which such an unqualified person ought to be regarded as sufficient where art 22 was at issue.

In the UK there is no authority on this matter but one particularly interesting example has been the Ofqual algorithm which was used in the summer of 2020

[17] See Bygrave L, commentary on art 22, in Kuner C et al eds The EU GDPR: A commentary (OUP, 2020), fn 37 .

to provide students who had been unable to sit their exams in the pandemic, with alternative estimated grades, partly based on teacher estimated rankings of students, and partly on the historic performance of their schools. Ofqual claimed that the results of students generated by the algorithm were not "solely automated decisions" because humans (teachers) determined one of the inputs (ranking of students at a particular school) and exam centres were reviewed before sign-off. However, this interpretation misconstrues the nature of human involvement required to bring a system outside the scope of article 22. **In almost all real-world cases, inputs to an algorithm will be to some extent produced by humans; this interpretation would therefore bring almost any conceivable algorithmic system out of the scope article 22,** including to the specific examples of ADM mentioned in Recital 71 of the GDPR (automated credit assessment and e-recruitment). Human production of inputs is not the same as human review of the output. In this case, the role of the human reviewer was merely to determine in what order the grades which were algorithmically rationed at the school cohort level would be distributed within that cohort; they had no power to award an A when the algorithm had already determined that nobody from their school cohort would get an A. Furthermore, the review of an exam centre was not the same as review of a decision relating to an individual student. In neither case does it appear that humans meaningfully reviewed the decision reached by the system relating to an individual pupil[18]. The matter became moot when the automated estimated ranking system collapsed under public opprobrium and perceptions of bias[19].

In the Amsterdam *Uber* and *Ola* cases, the question of what was a solely automated decision was partially addressed by a court. In these cases, the guidance of the A29 WP was strongly drawn upon. In the 11 March 2021 decision relating to *Uber* drivers,[20] the Court found that Uber's automated contract termination procedure did not constitute a SAD under Article 22. Uber argued that although a decision might be taken without human intervention to temporarily block access to the Driver app after a fraud signal was automatically generated, this was not the entirety of the decision to terminate the contract. Instead, Uber said, and the court accepted, termination was actually a decision by (at least) two employees of the Risk team on the basis of an investigation conducted in response to the fraud signals. The court thus accepted that although the decision to temporarily block access to the Driver app after a fraud signal was solely automated[21], the decision to terminate was not.

---

[18] See Ada Lovelace Institute, "Can algorithms ever make the grade?", 2020 at https://www.adalovelaceinstitute.org/blog/can-algorithms-ever-make-the-grade/#fnref-13

[19] See Office for Statistics Regulation Ensuring statistical models command public confidence Learning lessons from the approach to developing models for awarding grades in the UK in 2020 March 2021 at https://osr.statisticsauthority.gov.uk/wp-content/uploads/2021/03/Ensuring_statistical_models_command_public_confidence.pdf

[20] C / 13/692003 / HA RK 20-302 .

[21] And this initial decision was regarded as not having "significant" effects and thus did not qualify as an art 22 SAD - see below.

In the *Ola* decisions, the applicants similarly failed in most cases to prove the existence of a solely automated decision (SAD). For example an automated profiling system which gave drivers bonuses for certain parameters such as turnover, hours worked, etc was not deemed “significant” (below). However, the system which deducted pay as penalties for allegedly fraudulent behaviour by drivers, was held to be a SAD, but with little reasoning separable from the question of “significant effect” discussed below (para 4.51) [22]. Finally in another Amsterdam *Uber* decision decided by default on 24 February 2021 but not published till 14 April[23], five Uber drivers argued that they had been wrongfully accused of fraudulent activity and consequently dismissed by Uber by solely automated decision. Since Uber did not appear, the court upheld this, even though it seems to have involved the same contract termination system that was rejected as a SAD above[24].

3.4.1.2 Legal or “significant” effects

Secondly the decision must have legal or “significant” effects. The A29 WP suggested that “significant” decisions include those with the potential to “significantly influence the circumstances, behaviour or choices of the individuals concerned”, as well as those that may lead to individuals' “exclusion or discrimination”. This may not seem a difficult bar to reach for most public sector risk scoring or benefit allocation systems, but it might be more problematic with respect to individual building block systems that triage upstream decisions, or produce inputs to an eventual high stakes decision, especially when outsourced wholly or partly, explicitly or tacitly, to private operators.

The *Uber* court of 11 March, as noted above, accepted that the decision to temporarily block access to the Uber Driver app after a fraud signal was made without human intervention. However, this temporary blocking had no long-term or permanent effect, so the automated decision had no legal consequences and did not significantly affect the driver, and so was not deemed a SAD for art 22. In the *Ola* cases of 11 March, the court held similarly about apps or parts of apps which allocated bonuses to drivers for certain actions, flagged irregularities in driving behaviour, and matched drivers to passengers. All of these operated without human intervention, and all possibly had some influence on the driver's behaviour, but this was not deemed a legal or “significant” effect. Only the automated pay deduction system did reach this threshold. “[T]he decision to impose a discount or fine has effects that are important enough to merit attention and that significantly affect the behavior or choices of the person concerned as referred to in the Guidelines. After all, such a decision leads to a sanction that affects the rights of [applicants] under the agreement with Ola” (para 4.51). The existence of profiling alone was not enough to generate significant effects, since a number of the other app-based systems were deemed as profiling for the purpose of founding art 15 access rights.

---

[22] C / 13/689705 / HA RK 20-258.
[23] C/13/696010 / HA ZA 21-81
[24] n 25.

Through UK eyes, these cases seem rather unsatisfactory as precedents, involving partial submission of evidence by applicants, little detail in the decision and some undefended process by Uber. Nonetheless the default Uber judgment is a landmark in leading to what seems to be the first EU judicial reversal of an automated decision as a result of art 22, leading to the reinstatement of the automatically sacked driver.

Development of case law in the UK would help in reducing uncertainty in this area, as ICO and EDPB guidance though useful, remains simply that. Failing that however, more detailed statutory definitions of these terms in the post Brexit era, might help, as might sector-specific codes.

3.4.1.3 What is the "decision" that is solely automated?

Underlying these examples and cases is a basic problem: what exactly is the *decision* that is solely automated? When, if ever, is a decision that has automated and human elements "*solely automated*" and can we find general rules to decide on this matter? And in any case, is this really a helpful way of categorising decisions, especially in public sector high stakes decisions, as requiring special attention and safeguards?

An ADM system may be said to fall into one of three categories; *supporting* a human eg a decision support system such as a court sentencing assistant which provides guidance to a judge or official; *triaging* (determining, via an automated process, what cases or applications are passed wholly or partly to a human decision maker, or which cases are handled by which decision-makers) - eg an initial alert flagging possible fraud on a benefit system which leads to human attention); or *summarising* (using automated processing to summarise decisions from more than one human decision maker) - eg where human ratings are systematised into a single score as in many gig economy rating scenarios.[25]

In each of these, there are many situations in which it will be unclear whether and when a solely automated decision is made. In each, there is (at least) a solely automated part and a human part comprising a whole system leading to a decision or decisions which may have legal or significant effects. Depending on the output of the system (e.g. flagged as high risk or low risk), the level of scrutiny by the human decision-maker might be above or below the threshold of solely automated decision-making. Depending on output, the *effect* might be significant or not. Significance will be contextual but to which part of the context? Sometimes the human decision precedes the automated part; sometimes the reverse. There may be feedback and synergy. Especially if these different parts are distributed across different organisations and the private and public sector, it may be extremely hard to discern if a solely automated decision has indeed occurred which triggers art 22 rights. The issue is much wider than merely art 22: as we see below in 4.2.2.4 it is also relevant to judicial review, and what factors either a human or a system took into account.

[25] See Binns R and Veale M, (draft under review) "Is That Your Final Decision? Multi Stage Profiling and Selective Effects Under Art 22 of the GDPR"

It seems unlikely that hard or bright lines can be found to declare which of the many possible hybrid machine; human systems should be deemed a solely automated decision with legal or significant effects. **Taken together with the other nested uncertainties of the text, as Binns and Veale comment, the net conclusion here may be that art 22 is "conceptually beyond saving"[26].**

3.4.2 Lawful ground for SADs

Under art 22(2)(b), SADs are lawful for public bodies to take where they are "authorised by law" but only subject to certain safeguards discussed below. Examples given in recital 71 relevant to 22(2)(b) include fraud and tax evasion monitoring and "prevention purposes" for regulation.

Furthermore, article 22(4) provides that where SADs are based on Art 9 sensitive personal data (SPD), processing must be based on explicit consent or substantial public interest, with probably only the latter being appropriate to public ADM[27]. Note, as discussed above in section 3.3, that many public sector systems will "reveal" SPD. In that case, according to art 9(2)(g), processing must be proportionate to the aim pursued, respect the essence of the right to data protection and provide for suitable and specific measures to safeguard the fundamental rights and the interests of the data subject. Art 22 then makes further requirements for safeguards for both types of SADs.

3.4.2.1 Safeguards and explanations

Where art 22(2)(b) is relied on for SADs "authorised by law", "suitable measures to safeguard the data subject's rights and freedoms and legitimate interests" must be put in place (art 22(2)(3)) and the same applies to art 22(4) SPD SADs. Recital 71 states in relation to both SADs, that "suitable safeguards" "should include specific information to the data subject and the right to obtain human intervention, to express his or her point of view, to obtain an explanation of the decision reached after such assessment and to challenge the decision".

From these words a heated debate arose across the EU from 2016 on as to whether art 22 did indeed confer a "right to an explanation" concerning (some) algorithmic decision-making and if so, when and what this meant, even though no such right was explicitly found in the main text. It was hoped such a right

---

[26] Ibid. As this report was finalised,an independent report of the Taskforce on Innovation, Growth and Regulatory Reform (TIGRR) lead by 3 Tory backbench MPs (Ian Duncan-Smith, Theresa Villiers and George Freeman) was issued on 16 June 2021 which examined various areas for legislative change in the space opened up by exit from the EU and called for the abolition or minimisation of art 22 (pp 49-53). "Article 22 of GDPR should be removed. Instead a focus should be placed on whether automated profiling meets a legitimate or public interest test, with guidance on how to apply these tests and the principles of fairness, accountability and an appropriate level of transparency to automated decision-making provided by the Information Commissioner's Office"(para 225) . See https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/994125/FINAL_TIGRR_REPORT__1_.pdf .

[27] See n 12 above.

might expose and allow challenge to discrimination or bias in complex and opaque algorithms; although as noted in Appendix A, such "black box" issues are rarer than might be supposed in single-system public sector ADMs, and explanation rights might possibly more usefully be aimed at the provenance of training data and models in supply chains of providers and sub-providers, rather than the decision made by a single public data controller at the apex of the processing pile.

In the DPA 2018, s 14, the UK decided to implement the required safeguards under art 22(2)(b) as follows:

> (i) the controller must, as soon as reasonably practicable, notify the data subject in writing that a decision has been taken based solely on automated processing, and
>
> (ii) the data subject then has 1 month in which to ask the data controller to reconsider the decision, or take a new decision that is not based solely on automated processing.
>
> (iii) the data controller then has 1 month, or in complex cases the possibility of a further 2 months (GDPR art 12(3)) to "consider" the request, comply with it and notify in writing the steps taken to comply and the outcome of these steps.

The lack of any explicit reference to a "right to an explanation" in the DPA 2018 s 14 was raised during its passage through the Lords but did not produce any definite response, although the Secretary of State retains the possibility to review and make new regulations (s 14(7)). **It would be helpful to clarify in statute if such a right exists by virtue of art 22 in public sector ADM and if so, if and how it differs from the information and subject access rights discussed in 2.3 below.**

There are other potential gaps here also.

(i) *Does "reconsider" imply a data subject has the right to submit new evidence to a human why the SADM system is erroneous or unfair or can any new human decision lawfully be based only on the same information the automated system used?* The *Ladd v Marshall* principles[28] state that new evidence can be admitted when it fulfils three conditions: first that the evidence could not have been obtained with reasonable diligence for use at the trial; second, the evidence must be such that, if given, it would probably have an important influence on the result of the case, though it need not be decisive, and third, that the evidence must be such as is presumably to be believed, or in other words, it must be apparently credible, though not incontrovertible. The latter two issues will be a question of fact in each individual case, but given that the claimant may well not know what factors are being considered by an ADM system in the first instance, and given the lack of control the claimant has over what that system considers, it seems

[28] [1954] EWCA Civ 1.

very likely that the first of these will be satisfied. Certainly, if it can be satisfied when the first hearing was in a court, it seems even more likely that it can be satisfied when the first decision took the form of ADM.

(ii) *Does the "or" in (ii) above mean that reconsideration might be not by a human but by another automated system*? Would it suffice for the system to be the same but retrained? Does the user have the right to decide which option they would prefer?

(iii) *Can the data controller be forced, as part of "considering" the request, to ask data processors or joint controllers to reconsider decisions further down the stack?* If they cannot, can they really substantively reconsider their decision? If a decision upstream is found to be flawed then it seems at least likely that a decision dependent on it further downstream might also be flawed for, e.g. error of law in the case of a public authority.

3.5 Articles 13-15

Perhaps better alternate routes to a right to an explanation, if not by this name, exist via the well-established transparency rights of the GDPR in arts 13-15.

First, at the time the data are obtained from the data subject, Art 13(2)(f) gives the data subject a right to know the existence of the ADM including profiling; a right to 'meaningful information about the logic involved' and the right to know the significance and envisaged consequences of such processing for the data subject.' Similar rights are contained in Art 14(2)(g) which deals with when data is not obtained directly from the user. Both these provisions operate passively, requiring the data controller to give the relevant information to the data subject. Art 15 however gives the data subject the right to obtain this information actively and again art 15(1)(h) refers to "meaningful information about the logic of processing".[29]

Substantial debate exists about what exactly is meant by 'meaningful information about the logic involved', and the academic battles on this point are well ventilated. Goodman and Flaxman had initially argued that this provision conferred a right to a 'full explanation' which would limit the ability to use machine learning that acts in a more opaque manner.[30] Wachter, Mittelstadt and Floridi then countered that 'the right of access… only grants access to an ex ante explanation of system functionality… [I]t is reasonable to doubt that the right of

[29] Art 29 Working Party, now known as the European Data Protection Board (EDPB) 17/EN WP251rev.01 'Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679.'6/2/18 at 26.
[30] B Goodman and S Flaxman, 'EU regulations on algorithmic decision-making and a 'right to explanation" AI Magazine, Vol 38, No 3, 2017, arXiv:1606.08813

access grants a right to ex post explanations of specific decisions already reached.'[31]

More recently the debate has moved towards suggesting that the answer to this question is in fact flexible and context-specific.[32] Similarly, the Information Commissioner's Office (ICO) and Turing Institute's report, which identifies six types of explanation[33] concludes that 'when we talk about explanations… we do not refer to just one approach… or to providing a single type of information… Instead, the context affects which type of explanation you use to make an AI-assisted decision clear or easy for individuals to understand.[34]

The *Uber* and *Ola* decisions mentioned above are good examples of applicants managing to obtain some "meaningful information" about algorithmic management using art 15 (to its full extent - not just art 15(1)(h)). In relation to the deduction of earnings system, the *Ola* court found that 'Ola must communicate the main assessment criteria and their role in the automated decision to [the drivers], so that they can understand the criteria on the basis of which the decisions were taken and they are able to check the correctness and lawfulness of the data processing' (para 4.41). It seems here that an ex ante approach of the type favoured by Wachter et al was adopted.

However, other requests for information by the drivers made under 15(1) generally were rejected, mainly because the data pertained to other data subjects (passengers) as well as drivers, eg customer booking data, customer ratings of drivers, GPS data about where the cars were during the trip. The court has discretion to refuse access if it is "necessary for the rights and protections of others" (art 15(4)) although the court acknowledged this should be applied restrictively. But another problem was that the requests for data made by drivers were sometimes deemed lacking in detail, partly because of insufficient knowledge about the systems which *in itself* hampered them in knowing what to ask for.

Other problems exist with using these art 13-15 rights in the ADM context. We have seen that to trigger Art 22, the decision must be based *solely* on automated

---

[31] 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 at 17 and see also 19.

[32] A Selbst, J Powles, 'Meaningful information and the right to explanation' , 2017 *International Data Privacy Law* 233; M Kaminski, 'The Right to Explanation, Explained' [2019] 34 *Berkley Tech LJ* 189; Art 29 Working Group (now EDPB) Guidelines on Automated Individual Decision-Making and Profiling at 25-6.

[33] *Explaining decisions made with AI*, ICO and Alan Turing https://ico.org.uk/for-organisations/guide-to-data-protection/key-data-protection-themes/explaining-decisions-made-with-ai/ .The six types of explanation are the rationale (reasons leading to a decision); responsibility (who is involved and whom to contact); an explanation of what data has been used and how; an explanation of how equal and equitable treatment is to be ensured; an explanation of the system's safety and performance in terms of reliability, security and robustness; and finally an explanation of the potential impact of the system on individual and society.'

[34] Ibid at 21.

processing, Art 13-15 rights are generally triggered as long as there is *any* automated processing. However the key rights in arts 13-15 to "meaningful information about the logic involved" are restricted to "automated decision making, including profiling, referred to in art 22(1) and (4) .. at least in those cases..". There has been inclarity as to whether this means that art 15(1)(h) is in fact as limited in scope as art 22 by the requirements of "solely automated" and "significant"[35]. Unfortunately the *Ola* court of March 11 seems to have accepted this restrictive approach in para 4.49 : " The court therefore assumes that there is no automated decision-making within the meaning of Article 22 paragraph 1 GDPR. As Article 15 (1) (h) GDPR only applies to such decisions, this part of the request is rejected". This seems unfortunate.

Finally, there has always been an exception to DP transparency rights in the form of the protection of trade secrets and intellectual property. A request for the logic behind a private sector vendor of an outsourced ADM system or component might well run up against an IP defense. This again seems very significant in the light of a landscape of at least partially procured systems from private vendors. Recital 63 of the GDPR does however counsel that this should not justify "a refusal to provide all information to the data subject". Note this says nothing about transparency to the public operator: thus provision in procurement frameworks for access to data (user and training set) as well as logic (models, algorithms, rules) is likely to be crucial. In the UK this has not yet raised its head as a major problem at least in disputed cases, partly because major systems have tended to be built in house[36]; however in the US it is a well known issue[37].

It would be helpful if these uncertainties as to what "meaningful information" means, how it could/should be applied in practice, and in different contexts and what restrictions should be placed upon it, could be clarified.

3.6 Bias, discrimination and fairness in the GDPR; building better systems

The GDPR is almost entirely silent, on bias and discrimination in ADM. The most direct allusion to it is found in recital 71, which notes the controller should

---

[35] See Edwards and Veale, p24-25 where it is argued this is the wrong conclusion given the words "and, at least".

[36] Interestingly however the UK government has protected the source code of UK AI products from access by Japanese authorities under the UK-Japan Comprehensive Economic Partnership Agreement (CEPA) : see https://questions-statements.parliament.uk/written-questions/detail/2020-11-12/114874 . See also from Sweden,Christensen, Kristina LU *Exhibiting transparency without opening the 'Black Box' - Balancing act between Data Protection and Trade Secrets Rights in Solely Automated Decision-Making AI system in Healthcare* (2020) JAEM03 20201 at https://lup.lub.lu.se/student-papers/search/publication/9019754 .

[37] Diakopoulos, Nicholas. 2014. *Algorithmic Accountability Reporting: On the Investigation of Black Boxes*. Columbia Journalism School: Tow Center for Digital Journalism.

"implement technical and organisational measures […] in a manner […] that prevents, inter alia, discriminatory effects on natural persons" on the basis of special categories of data. This can be read as promoting but not mandating "discrimination aware" or "fairness aware" data mining and machine learning, a growing field of research and practice[38]. While "fairness" *is* a mandatory principle of the GDPR (art 5(1)(a), it is an extremely "elusive" notion in data protection[39]. Controls over bias and fairness in the context of public ADM are probably far more accessible via judicial review and equality law, discussed below.

It is important to note that it is not, in the main, the job of the DP regime to prescribe how a better, less biased, more inclusive ADM system might be built. Data subjects are at least in theory given active rights to transparency, to rectification, to erasure, to object to profiling and more; **but there is no active right to shape *how* inferences are made by ML automated systems, and *what* they are.** Wachter et al have argued that individuals are thus in practice granted little control and oversight over how their personal data is used to draw inferences about them and have suggested a "right to reasonable inferences" at least in certain scenarios. This right would require ex ante justifications to be given by the data controller to establish whether an inference is reasonable. This disclosure would address (1) why certain data form a normatively acceptable basis from which to draw inferences; (2) why these inferences are relevant and normatively acceptable for the chosen processing purpose or type of automated decision; and (3) whether the data and methods used to draw the inferences are accurate and statistically reliable[40].

To apply such aspirational ideas to the public sector is a huge job which is perhaps not best commenced by starting from reconsidering DP law. Instead it might be better to think, as we suggest below (section 5), of looking at what new, constitutive regimes might be devised to regulate automated systems in the public sector or, as with the template model of the proposed EU AI regulation, in the round across private and public sectors. Indeed, as will be discussed further below, if public law is better able to develop its 'public wrongs' approach in order to control and optimise ADM in the public sector, there may well be elements of this approach that could be extended into the private sector, without the need to rely on changes to private rights.

---

[38] A29 WP Guidance suggests measures to tackle discrimination under Recital 71, including that data controllers "design ways to address any prejudicial elements", "audit algorithms", and undertake "regular" and "cyclical" reviews to avoid discrimination on the basis of SPD. Worth noting also that the UK in DPA 2018 Schedule 1, 8.1(b) specifically provides a lawful basis for processing sensitive personal data expressly to "debias" machine learning systems (as an aspect of substantial public interest).

[39] See further Clifford, D and Ausloos, J "Data Protection and the Role of Fairness" 2018 37 Yearbook of European Law130–187, https://doi.org/10.1093/yel/yey004

[40] Wachter, Sandra and Mittelstadt, Brent, A Right to Reasonable Inferences: Re-Thinking Data Protection Law in the Age of Big Data and AI (October 5, 2018). Columbia Business Law Review, 2019(2)

3.6.1 DPIAs

One part of the GDPR which is expressly designed to help proactively build better systems is article 35 which requires a data controller to conduct a data protection impact assessment (DPIA) where the processing is likely to result in high risk to the rights and freedoms of natural persons. Art 35(3) (a) requires a DPIA where in particular there is a:

> "systematic and extensive evaluation of personal aspects relating to natural persons […] based on automated processing, including profiling […] and on which decisions are based that produce legal effects concerning the natural person or similarly significantly affect the natural person."

This will very often be the case when a public ADM system is constructed and indeed both EDPB[41] and ICO guidance point towards considering a DPIA in almost all plausible high-stakes public sector ADM scenarios (NB. not necessarily *solely* automated). Such DPIAs are extremely useful for all concerned. They allow the data controller to document that they have assessed, thought about and mitigated potential privacy risks in advance. They allow users at least in theory a chance to participate in shaping the system. And they have become extremely useful de facto tools with which civil society can analyse and critique new high risk systems**. It would be very useful for the role of DPIAs to be expanded and clarified[42].** Some ways to do this would be:

A. to provide explicitly that DPIAs be compulsory for all public sector ADM systems
B. to demand publication, within a certain period, possibly with redactions for sensitive or IP related material
C. to mandate certain types of user consultation (this is currently quite restricted by art 35(9), where consultation is required only 'where appropriate').

Publication within a reasonable period, ideally before the system starts to operate, has been a constant running sore in respect of key public sector systems in recent years. Publication can sometimes be mandated or incentivised by an FOI request although exceptions can make this a protracted process. It would be better for public confidence as well as scrutiny if publication was required for all but the most sensitive public sector ADM systems.

3.7 Conclusions about the DP regime

As the above shows, the DP regime, despite its prominence in the literature, is not necessarily the optimum place to regulate ADM systems in the public sector. Issues canvassed above show a large number of gaps. The DP regime is

[41] A29 WP Guidelines on Data Protection Impact Assessment (DPIA) and determining whether processing is "likely to result in a high risk" for the purposes of Regulation 2016/679 | WP 248 rev.01 (13 October 2017) https://www.dataguidance.com/sites/default/files/wp29-gdpr-dpia-guidance_final.pdf

[42] See also Swee Leng Harris "Data Protection Impact Assessments as rule of law governance mechanisms" Data & Policy (2020), 2: e2, 1–21 doi:10.1017/dap.2020.3 .

strongest in providing *procedural, transparency rights* in arts 13-15 which can then be used to found challenges by users (as in the Uber/Ola cases) but even those may be very restricted by factors including scope, trade secrets and practical issues of how to convey meaningful information relating to systems or individual decisions. None of these are helped by the absence of UK or CJEU case law.

DP in the context of public sector ADM largely does not give users practical rights to control the *substance* of decisions taken although the underdeveloped fairness principles in art 5 might be of some use. Art 22(2)(b) does allow for routes to challenge a SAD, enabled for the UK by DPA 2018, s 14, but not the far more common *assisted* automated decision.

However on a more positive note, DPIAs are already a useful tool for scrutiny of and to incentivise better prior planning of fairer and better ADM systems, and could usefully have their scope expanded, possibly in combination with public sector Equality Impact Assessments (EIAs), see next section.

4. Current legal framework: Equality; judicial review; public laws; and procurement

4.1 The Public Sector Equality Duty (PSED) and Human Rights Act (HRA) 1998

The Equality Act 2010 149(1) creates the PSED; the duty of public authorities to have due regard to the need to eliminate discrimination, harrassment, vicitimisation and any other conduct prohibited by the 2010 Act and both advance equality of opportunity and foster good relations between those who share a relevant protected characteristic and those who do not. S 1(3) explains that this involves positive steps to minimise disadvantages and to take steps to meet the different needs of those with protected characteristics. These characteristics are, of course, age, disability, gender reassignment, pregnancy and maternity, race, religion or belief, sex and sexual orientation. The HRA 1998 also applies to ADM systems where they impinge on one or more of the human rights listed in the European Convention on Human Rights (ECHR)[43].

It was these two pieces of legislation which played a central role in one of the first instances of judicial review to tackle the challenge of public authority use of ADM, namely *Bridges,*[44] concerning South Wales Police (SWP)'s use of Automated Facial Recognition technology (AFR). The Court of Appeal held that

[43] For further discussion on the equality implications of AI and ADM read AI Law Consultancy's opinion for TLEF: In The Matter Of Automated Data Processing In Government Decision Making

[44] *R(Bridges) v Chief Constable of South Wales Police* [2020] EWCA Civ 1058.

the use of the technology was not 'in accordance with the law' for the purposes of Art 8(2) of the ECHR, because there was no clear guidance on where the technology could be used and who could be put on a watchlist.

The Court also found that SWP had not done all they could reasonably do to discharge the PSED because it was a privately-developed system, as discussed above, SWP had not had access to the system's training data to assess its democratic composition and had thus had no means of knowing whether or not it was biased. And second, although the data of those who did not trigger the system was not retained for data protection purposes, this also left SWP without a means of assessing the false negative rate of the AFR. SWP had argued that the data of those who did generate an alert did not indicate any evidence of bias, but this is to make the mistake outlined above. Just because the overall results of the system appeared to match society generally this did not mean that the system was performing acceptably against more specific metrics. And in any event the court also emphasised the procedural nature of the PSED. Even if the output of the AFR had been acceptable, that did not mean that the process of deploying it had complied with the PSED. The challenge, of course, is to know how public authorities can, in the light of *Bridges*, ensure that their systems are compliant with the PSED. **We have already noted the difficulties involved in getting further access to the system and its data as a result of the inevitable IP concerns involved when the system is privately developed. So in the absence of such access it will be necessary to develop other systems for ensuring that privately developed systems are compatible with the PSED before they are deployed. At present it is not clear how this might be done, though kitemarking is one possibility.** [45]

4.2 The Common Law of Judicial Review

It is well known that judicial review acts as a referee[46] but also as a 'judge over the shoulder',[47] providing ex ante guidance to public authorities as well as an ex post challenge. This means that judicial review need not be, and indeed should not be, wholly antithetical to the interests of public authorities. Enhancing the lawfulness of decisions made by public authorities can increase public trust in those decisions, something the ICO and Alan Turing Institute have noted particularly in the context of ADM.[48] It is likely that in future the principles of judicial review will be used to constrain and govern choices made about where, when, and how to deploy ADM systems.[49] The orthodox division of grounds of judicial review is the tripartite system of Lord Diplock in *GCHQ*, namely grounds

[45] See further R Williams, 'Rethinking Administrative Law for Algorithmic Decision Making', *Oxford Journal of Legal Studies*, forthcoming.
[46] *R v Secretary of State for the Environment ex p. Hammersmith and Fulham LBC* [1991] 1 AC 521 at 561, per Lord Donaldson.
[47] https://www.gov.uk/government/publications/judge-over-your-shoulder
[48] https://ico.org.uk/for-organisations/guide-to-data-protection/key-data-protection-themes/explaining-decisions-made-with-ai/ at 57.
[49] See further R Williams, above n 47.

of illegality, irrationality and procedural impropriety.[50] However, it may be more helpful to consider the grounds of review chronologically. On this basis a decision maker must

- Have jurisdiction to make a decision in the first place.
- Adopt the correct procedure and procedural protections in making the decision.
- Not fetter or delegate its discretion beyond the permitted extent.
- Take into account the right considerations and only the right considerations.
- Make a decision which is reasonable or proportionate (as appropriate).

In the context of ADMs there are two types of decision which are likely to be subject to review on the basis of these grounds; the decision to commission and deploy an ADM system in the first place and any decision which is subsequently produced by that system. It is also worth noting that while we are used to the use of ADM systems reducing the transparency of the decision making process (discussed in further detail below), it is also the case that the specification of rules in a rules-based system, or the labelling of training data or selection of features may actually render more transparent and more open (and thus more reviewable) decisions that might take place at a more subconscious level in human decision making. **While some grounds of judicial review are sufficiently developed to enable this to take place (and indeed here the challenge will be in ensuring those grounds are fully used), others will need considerably more development in order to apply effectively to ADM.**

Tables 1 and 2 below summarise the application of judicial review to ADM from two perspectives. Table 1 presents the intersection of grounds of judicial review with stages of decision-making and issues raised as a result of an increase or decrease in transparency resulting from the use of ADM systems. Table 2 presents the intersection of technical issues with ADM systems, and the possible grounds of review on which the ADM system could be challenged. The grounds for review in Table 2 are coloured according to a traffic light system, from those which are most readily applicable (green), where the challenge is for courts to realise the potential of the existing grounds of review, to those requiring significant work before they could be applied (red).

Table 1: Intersection of grounds of judicial review with stages of decision making and issues raised by increases or decreases in transparency

Key:
Arises from the greater transparency of ADM systems
Arises from the lack of transparency in ADM systems
Not related to issues of transparency

	Decision to deploy the	**Decisions made by or with**

[50] *Council of Civil Service Unions v Minister for the Civil Service [*1984] UKHL 9, [1985] AC 374.

	system	the system
Rules-based	•Who made decisions about the rules (delegation) •What are the rules? (Jurisdiction, relevance of considerations) •Was deployment of the system proportionate/reasonable? (proportionality, *Wednesbury*)	•Why did the system think a precedent condition was fulfilled? (jurisdiction) •Potential Rigidity of DM (fettering) •Potential over-reliance on ADM (delegation)
Machine Learning	•How and by whom was the system trained? (delegation, jurisdiction, relevance of considerations) •What metrics are being used to assess the reasonableness etc of its deployment? •Why was that system chosen as opposed to a different one? (fairness of procedures) •Was deployment of the system proportionate/reasonable? (proportionality, *Wednesbury*)	•Lack of transparency: GDPR Arts 13(2)(f), 14(2)(g), 15(1)(h) – meaningful information about the logic involved, notice/reasons, cf gisting at common law. •Potential over-reliance on ADM (delegation) •What metrics are used to assess its conclusions? (Jurisdiction, relevance of considerations, proportionality, *Wednesbury* etc).

Table 2: Traffic light notation of the intersection of grounds of review with the technical challenges arising from ADM

Technical issue	**Potentially applicable ground of review**
Outsourcing	Delegation
Metrics to be used in choosing/assessing systems	*Wednesbury* reasonableness, relevance/irrelevance, jurisdictional review,
Testing v deployment differences	*Wednesbury* reasonableness relevance/irrelevance
Discrimination and equality issues in data	Proportionality (necessity) Duty to take into account a relevant consideration
Population level results applied at individual level	relevance/irrelevance Fettering

Unobserved labels	Duty to take into account a relevant consideration
Opacity/lack of transparency	Procedural fairness (choice of procedure) Reason giving
Automation bias	Delegation
Rigidity	Fettering

Key:

- Ground exists but a significant amount of work is needed if it is to perform the necessary function in relation to ADM.
- Ground exists and is reasonably capable of performing the function needed in relation to ADM.
- Ground exists and is well developed in order to deal with ADM, but the challenge here is to ensure that the potential of the ground is recognised and understood, and that good use is made of the ground.

The content of these tables is then discussed in more detail in the next sections which will consider the following legal issues:

- Reviewing the process of commissioning and deploying an ADM system
    - The reasonableness or proportionality of adopting a particular system in a particular context.
    - The procedural fairness of using a particular system in a particular context.
    - The restrictions on delegating decisions in the process of designing the system.

- Reviewing a decision made by an ADM system
    - Decisions which establish whether a condition precedent is fulfilled for the purposes of determining jurisdiction.
    - The extent to which the system fetters the use of discretion, renders the decision-making too rigid, or enables too great a delegation of power.
    - The transparency (or lack thereof) of the system and the duty to give notice/reasons.
    - The relevance or propriety of factors taken into account by
        - A human taking the output of an ADM system into account
        - The ADM system itself.
    - The rationality or proportionality of the decision made by the system.

4.2.1 Reviewing the process of commissioning and deploying an ADM system

As noted above, in order to commission or deploy an ADM system at all, a public authority must, in the light of Art 22(b) GDPR, be able to point to a specific legislative basis for doing so.[51] However, such provisions tend to be permissive (stating that decisions *may* be made by a computer) rather than compulsory, meaning that the common law of judicial review has a role to play in establishing whether their use was in fact lawful in the particular circumstances.

4.2.1.1 The reasonableness or proportionality of adopting a particular system in a particular context.

In deciding whether or not a public authority 'may' adopt an ADM system at all, it is in fact the last of the grounds of judicial review which seems most applicable. This requires the public authority to consider whether deployment of an ADM system at all is reasonable[52] or proportionate,[53] in the sense that adoption of such a system must not be a sledgehammer to crack a nut. For example, where it is possible to adopt a form of augmented decision-making, rather than full automation, this is preferable, though as noted above, and discussed further below, this assumes genuine input from the human decision maker in order for the distinction between augmentation and automation to be meaningful. There may well also be a 'fair balance' issue relating to proportionality in the sense that it will be necessary to examine the benefits of the ADM against its impacts and potential disadvantages, such as the potential lack of transparency, or its potentially detrimental effect on minority classes of people. This latter point can be relevant even outside the PSED, as the ICO and Alan Turing Institute have also indicated.[54] From the point of view of reasonableness it will be here that the different metrics outlined above will really matter. There may well be hard choices to make between systems which are sufficiently accurate overall but perhaps have high false negative rates (what data scientists would call high precision, low recall), or which are sensitive enough to detect the target cases but also have a high level of false positives (low precision, high recall). What balance between these metrics will make it reasonable to use a system in which contexts? **At present Administrative Law has neither the technical expertise nor a developed enough account of what makes a decision reasonable in principle to be able to provide the necessary answers.**

---

[51] Such as, for example, Social Security Act 1998 S 2, or Child Support Act 1991 S 50A.
[52] *Associated Provincial Picture Houses v. Wednesbury Corporation* [1948] 1 KB 223 (CA).
[53] See, e.g. *Bank Mellat v Her Majesty's Treasury* [2013] UKSC 39.
[54] https://ico.org.uk/for-organisations/guide-to-data-protection/key-data-protection-themes/explaining-decisions-made-with-ai/ at 41.

4.2.1.2 The procedural fairness of using a particular system in a particular context.

If it is decided that some kind of system will be introduced, there will then be a series of decisions to be made about the kind of system to be deployed.[55] In addition, it is clear that the rules of procedural fairness in administrative law have not only been about ensuring that a given decision-making procedure is fair on its own terms but also about controlling the very choice of the particular decision-making system in the first place, even in the analogue world.[56] It therefore seems very likely that courts in future will be called upon to review the choice of one particular form of ADM as opposed to another. As noted above, there are many different forms of ADM, with, for example, different levels of explainability, and the ICO and Alan Turing guidance in particular is that 'the model you choose should be at the right level of interpretability for your use case and for the impact it will have on the decision recipient.'[57] **It is thus going to be necessary for courts to develop both the technical expertise to understand the differences between the different systems and a justifiable set of principles for deciding which system is most appropriate in which case.** This will almost certainly entail a need to move away from the courts' current general sense that the more court-like a procedure looks, the more fair it is,[58] since this is not nuanced enough an approach to harness the potential benefits of ADM systems as well as ensuring their fair use. There is a considerable amount of work to be done if the courts are to be able to answer this question effectively, and that work will involve the development of a greater technical understanding.

Once a system has been chosen, a set of further decisions will have to be made regarding its operation. For rules-based systems it will be necessary to establish precisely what those rules should be. For a machine learning system decisions will be made about the data and method by which it is trained. These choices have the potential to lead to judicial review, both of the choices themselves (discussed in the next section) and also the question of who makes them.

4.2.1.3 The restrictions on delegating decisions in the process of designing the system.

There are clear rules in administrative law which establish that while delegation is possible, not all decisions can be delegated.[59] This raises the difficult question

---

[55] It is important that these decisions are made by the public authority itself, as it seems likely that they would otherwise fall foul of the rules on the ability of public bodies to delegate their powers.*R v Adams* [2020] UKSC 19.

[56] See, for example the specification that a tribunal must be independent and impartial, under Article 6, rather than simply justifiable in its own right as a political decision-making method. (*R. v. Amber Valley DC, ex parte Jackson* [1985] 1 WLR 298, [1984] 3 All ER 501; *Ex p Kirkstall Valley* [1996] 3 All ER 304).

[57] See *Explaining Decisions made with AI* https://ico.org.uk/for-organisations/guide-to-data-protection/key-data-protection-themes/explaining-decisions-made-with-ai/ at 48.

[58] See above n 32.

[59] *Carltona v Commissioners of Works* [1943] 2 All ER 560; *R (CCWMP) v Birmingham Justices* [2002] EWHC 1087; *R v Adams* [2020] UKSC 19.

of where the line between the two should be drawn, and this line is particularly challenging in the context of ADM systems. It might be straightforward to establish that the main decision tree in a rules-based system, or the key features on which a ML system is trained should be decided by the public authority commissioning the system. But it is all too possible that what are thought of as more minor, implementational or engineering questions further down the line in fact turn out to have a significant impact on the operation of the system, meaning that there would be a loss of accountability if those decisions are taken by technicians in a private company supplying the system. For instance, without specific instruction, a data scientist might assume that different kinds of errors (e.g. false positives and false negatives) are equally problematic; or they might assume that there is no harm in aggregating two labels into one (e.g. aggregating 'Catholic' and 'Protestant' into 'Christian' for the purposes of discrimination analysis, which may or may not be warranted depending on the national context). It will be important for the administrative rules relating to delegation to be alert to this danger and capable of understanding and detecting how and when it arises in order to prevent this loss of accountability.

4.2.2 Reviewing a decision made by an ADM system

Once an ADM system is in place, the decisions it makes or assists will themselves be subject to review and it is here that some particularly challenging issues will arise. We are used to the idea that the 'black box' nature of some ML systems might pose problems as outlined above, but in some instances we also have almost the opposite problem. The process of specifying in detail the whole decision tree for a machine learning system, or specifying the relevant features on which a ML system should be trained, or the input variables in a regression model might in fact be *more* transparent than the equivalent processes in a human context, meaning that there is potentially a greater surface area on which administrative law can bite. And yet while **administrative law does in principle have grounds capable of dealing with these questions, it is very likely that those grounds are not yet nuanced and sophisticated enough to be able to do so in practice.** They may not be sufficiently developed on their own terms, there may be a lacuna in terms of the technical expertise necessary to apply them to ADM systems, or both.

4.2.2.1 Issue 1 - Determining jurisdiction

For example, as is well known, where a decision-maker's jurisdiction to take a particular decision is premised on certain factors ('if X1, X2, X3 are present, you may/shall do Y')[60] one of the first things a decision-maker will have to do is to determine whether these conditions are satisfied. But if this decision is made by an ADM system, reviewing it may well involve reviewing the specific criteria laid out for the determination in a rules-based system, or the labelling of training data in a ML one.

[60] See P Craig, *Administrative Law*, 8th ed (2016, London, Sweet & Maxwell) 16-001.

In addition, the decision in *South Yorkshire Transport*[61] suggests that in some cases at least the court will specify the definition of an 'X' condition to within a reasonable range and then leave the public authority to make its own decision within that range. But here again we meet the challenge of which of the different metrics listed above render such a decision 'reasonable'. How are we to choose which metric is appropriate for determining reasonableness in different contexts? In *Bridges*[62] one of the reasons why SWP did not fulfil the PSED was because they had not examined the false negative rate of the system used, with the resulting possibility that this false negative rate might be uneven across subgroups. It seems likely, therefore, that in future courts will hear arguments that ADM systems have used the wrong metric in determining the fulfilment or otherwise of an X condition in a statute. Here again it will be vital for courts to have a detailed understanding of the necessary technical arguments as well as the legal doctrine if they are to provide helpful guidance for those operating ADM systems and protect those who are subject to them. **It is also clear that the current law is not sufficiently developed to provide this guidance and protection.**

4.2.2.2 Issue 2 - Fettering, rigidity and over-delegation

However the jurisdiction of a public authority is determined, whether by human input or through an ADM system, once that public authority has jurisdiction it must go on to make the decision. Even in a wholly human context there is the potential for regular, routine decision-making to be guided by policies or even algorithms,[63] but in order to ensure that this does not wholly prevent decision-makers from exercising their discretion in the individual case, the law has developed the doctrine of non-fettering in cases such as *British Oxygen*[64] and *Stringer*.[65] And in addition, as noted above, the courts have developed rules which prevent public authorities, themselves the delegates of power from Parliament, further delegating those powers beyond what is strictly necessary for the purposes of efficiency.[66] **This suggests that it will be difficult for decision-makers to justify relying completely on algorithms. Exclusive reliance could be provided for by legislation of the kind noted above,[67] but the need under s 22(b) of the GDPR to ensure that such legislation contains 'suitable measures to safeguard the data subject's rights, freedoms and legitimate interests' might still pose difficulties.**

However, even where the decision is nominally that of a human decision-maker we have noted (section 3.4.1.1) the potential for over-reliance on and over-confidence in ADM systems. It will be important, therefore, for the process

---

[61] *R v Monopolies and Mergers Commission ex p South Yorkshire Transport* [1993] 1 WLR 23 (HL).
[62] *R (Bridges) v Chief Constable of South Wales Police* [2020] EWCA Civ 1058.
[63] See, e.g. *R(Guittard) v Secretary of State for Justice* [2009] EWHC 2951.
[64] *British Oxygen v Minister of Technology* [1971] AC 610.
[65] *Stringer* [1970] 1 WLR 1281
[66] *Carltona v Commissioners of Works* [1943] 2 All ER 560; *R (CCWMP) v Birmingham Justices* [2002] EWHC 1087; *R v Adams* [2020] UKSC 19.
[67] Such as, for example, Social Security Act 1998 S 2, or Child Support Act 1991 S 50A.

of judicial review to examine such systems closely to ensure that they are not used too rigidly (perhaps particularly in relation to rules-based systems) and to ensure that if there is a 'human failsafe' that person has a sufficient understanding of the operation of the system and, connectedly, does not rely too closely on it. In this respect the decision in *Bridges* is encouraging, holding that the 'human failsafe' was (a) not sufficient at the end of the decision-making process to supply the compliance with the PSED that had been necessary throughout, (b) likely themselves to be fallible, particularly in the context of identification and (c) that the relevant checks were not technically sufficient or (as noted above) carried out by an expert in data science. Similarly, the US case of *Loomis*[68] established that courts must explain the factors *in addition to* the algorithmic risk assessment system COMPAS that had been used independently to support the sentence imposed by the court. 'A COMPAS risk assessment', held the court, 'is only one of many factors that may be considered and weighed at sentencing.'

As far as fettering is concerned, this ground of review may well be relevant to the issue of individualisation of decisions, discussed above. It is well known that 'big data' can lead to more tailored solutions in advertising or medicine, but unless this micro targeting is as accurate as case by case individual decision-making it is still distinct from true individualisation. **If the doctrine of non-fettering does not do sufficient work here, we run the risk that once an ADM system is in place it will pull decision-makers towards an approach in which the efficiency and speed of routine and automated decision-making are prioritised, even when this leads to injustice or unfairness in individual cases.** And of course there is a high risk that those most likely to suffer these individual injustices are those most likely to be already at a disadvantage, as a result of the inequalities baked into the systems through the use of data.

4.2.2.3 Issue 3 - Challenging the decision itself - transparency

Even if the decision to deploy the system is legal, decisions made by that system can still be challenged.

It is here that the issue of transparency or lack of it may arise. This can give rise to judicial review as a matter of common law on its own terms, but of course it may also inform what is meant by 'meaningful information about the logic involved', as required by Articles 13(2)(f), 14(2)(g) and 15(1)(h) of the GDPR, at least when those Articles are applied to public authorities. Two sets of rules deal with these issues at common law: those regarding the duty to give reasons (if any) after the decision[69] and those specifying the claimant's right to have notice of the case against them for the purposes of a fair hearing.[70] These two sets of

---

[68] *State of Wisconsin v Loomis* [2016] WI 682016 WI 68, 371 Wis 2d 235, 881 NW 2d 749.

[69] See e.g. *R v UFC, ex p Institute of Dental Surgery* [1994] 1 WLR 242; M Elliott, 'Has the Common Law Duty to Give Reasons Come of Age Yet?' [2011] Public Law 56.

[70] See e.g. *Secretary of State for the Home Department v AF (No 3)* [2009] UKHL 28, [2010] 2 AC 269; following *A v United Kingdom* (2009) 49 EHRR 625 (GC).

rules are often connected in the sense that reasons given after one decision are often instrumentally relevant to any subsequent appeal, at which point they become more akin to notice. This is also reinforced by the way in which they are treated by the ICO guidance which states that understanding the reasoning behind a decision is 'vital' because it 'allows [applicants] to assess whether they believe the reasoning of the decision is flawed.' 'Knowing the reasoning supports them to formulate a coherent argument for why they think this is the case.'[71]

The duty to give notice has received particular attention recently in the context of closed material procedures (CMPs) where it has been held that the defendant must often be told the 'gist' of the case against them.[72] In *Bourgass*[73] Lord Reed held that:

a prisoner's right to make representations is largely valueless unless he knows the substance of the case being advanced. . . .… That will not normally require the disclosure of the primary evidence . . .… [but] what is required is genuine and meaningful disclosure of the reasons why [the decision was made].[74]

General statements about the prisoner's behaviour or risk were not held to be sufficient for this purpose. **All this tends to suggest that if the law would normally require 'gisting' (such as where, for example, the claimant's liberty was at stake), it should require at least the same level of information from an ADM system,** and indeed the technical arguments against reason-giving might be thought not to outweigh the need for such 'gisting' even outside the context of liberty and security of the person. In this context, then, the problem is not that the law does not have a solution to the issue, quite the contrary. The challenge will be in ensuring that the work done on 'gisting' is understood and used in the context of ADM decision-making. Some hope that this might happen comes from the US case *Houston Federation of Teachers v Houston Independent School District*[75] (HISD) which concerned the use by HISD of Educational Value-Added Assessment System (EVAAS) scores generated by a privately developed algorithm. The teachers argued that they had no meaningful way to ensure that their scores had been calculated correctly, even though these scores could be used to terminate their contracts for ineffective performance. In response to this claim HISD asked for summary judgment against the teachers and this was denied on the basis that 'HISD teachers have no meaningful way to ensure correct calculation of their EVAAS scores, and as a result are unfairly subject to mistaken deprivation of constitutionally protected

---

[71] ICO and the Alan Turing Institute, 'Explaining decisions made with AI' (20 May 2020) <https://ico.org.uk/for-organisations/guide-to-data-protection/key-data-protection-themes/explaining-decisions-made-with-ai/ at 23.

[72] See e.g. *Secretary of State for the Home Department v AF (No 3)* [2009] UKHL 28, [2010] 2 AC 269; following *A v United Kingdom* (2009) 49 EHRR 625 (GC).

[73] *R (Bourgass) v Secretary of State for Justice* [2015] UKSC 54.

[74] Ibid at [100].

[75] *Houston Federation of Teachers v Houston Independent School District* 251 F .Supp .3d 1168 (SD Tex 2017).

property interests in their jobs.'[76] Once the claimants had thus been permitted to proceed to trial, the case settled and HISD stopped using the EVAAS scores.

In France, legislation has gone even further, requiring users of ADM to release their source code, in addition to which the Digital Republic Law requires the publication of any source code used by government. In a rules-based system this is likely to be very helpful, while in a machine learning context it may well be much less so. As one of the authors has pointed out in previous work,[77] it is important that transparency is not seen as an end in itself or a panacea for all the ills that ADM might generate, but rather that the form of reason-giving that is chosen and required by judicial review is appropriate and useful in supporting other forms of accountability.

4.2.2.4 Issue 4 - Challenging the decision itself- factors taken into account

Transparency of reasons is vital in part because if we know what factors a decision-maker took into account, those choices can be challenged in line with the rules requiring decision-makers to take into account relevant considerations,[78] leave to one side irrelevant considerations[79] and avoid acting for improper purposes.[80] This ground of review has the potential to be very relevant to review of ADM, especially in the context of what is usually referred to as 'bias' (or as a lawyer would put it, discrimination or inequality of treatment). After all, *Wednesbury* itself gave the famous example of reliance on an impermissible, discriminatory characteristic (red hair) as something that it would prohibit.[81]

This ground is likely to arise at two potential stages. One is where a human decision-maker has to establish how much relevance to accord to the result of a determination by an ADM system, and second when the ADM system itself has taken various factors into account in producing a determination, either to give to a human or to execute on its own. **The law therefore needs a successful, clear and certain account of what is relevant and what is not in different circumstances, but we are in fact a long way from this.** Thus, for example, it is sometimes possible for decision-makers to take into account financial cost as a relevant factor in making a decision,[82] while at other times that is impermissible.[83] But the distinction between these cases and the answers given in them is by no means sharp or clear. If the law is thus currently unable to deal with a relatively familiar factor such as economic cost, it is to be expected that it will face even more challenges as it deals with ADM.

---

[76] Ibid.
[77] L Edwards and M Veale, 'Slave to the Algorithm? Why a 'Right to an Explanation' is Probably Not the Remedy You are Looking For' (2016) *Duke Law & Technology Review* 18.
[78] *Tesco Stores v Secretary of State for the Environment* [1995] 1 WLR 759.
[79] *R v Rochdale MBC ex p Cromer Ring Mill* [1982] 3 All RR 761.
[80] *R v Westminster Corporation v LNWR* [1905] AC 426.
[81] *Associated Provincial Picture Houses v. Wednesbury Corporation* [1948] 1 KB 223.
[82] *Ex p Barry* [1997] AC 584.
[83] *Tandy* [1998] AC 714.

Reviewing a human decision-maker's choice to take into account the output of an ADM system

If, for example, a human decision-maker takes the output of an ADM system into account, then whether or not that output was indeed relevant may well turn on its quality, particularly in the case of ML systems. But as noted above, there are a number of different metrics by which that quality might be measured, and not all of these will be consistent with each other. For example, a system might be fairly accurate overall but within that have a high level of false positives, or false negatives. A claimant might therefore wish to argue that by one metric the system was not performing well enough for its outputs to be taken into account in the decision-making process, while a defendant public authority might counter that by another metric it was. **If the law currently struggles to establish when financial considerations are relevant to a decision, it certainly has a long way to go in establishing which metrics render an ADM output relevant or not.**

Once a factor is regarded as being relevant, the precise weight to be given to it is currently for the decision-maker to choose.[84] But in the context of the PSED it is evident that this is not the case for some considerations relating to equality of certain protected categories. This is something we may well want to extend in the common law. If, for example, it is apparent that an ADM system is having a problematic effect in relation to a certain category of people, then even if they are not protected by the PSED[85] we might want this to be a relevant factor that the decision-maker *should* take into account in assessing the performance and thus the relevance of the ADM output. Similarly, we might regard it as highly relevant that, as noted above, the system was tested in a different environment from the one in which it is now being deployed. Or that the training of the system suffered from what we have referred to as 'unobserved labels'; the scenarios about which we do not have data, such as the higher risk defendants who were given community sentences. In other words, on all these kinds of issue we might want to encourage the courts to take a less deferential approach than they currently do under the *Tesco* doctrine,[86] and encourage public authorities to ensure that these matters are considered in deciding whether or not to base a decision on an ADM output.

Reviewing the factors that the ADM system itself 'took into account'

If the challenge is not to the fact that a human has taken an ADM output into account, but rather to the factors 'taken into account' by that ADM system itself, a further set of issues arise. **ML does not in fact reason through to decisions, it makes statistical inferences, operating on the basis of correlation, not causation. And, as observed at the outset, this can mean that while a system is accurate at population level, it can be much less so at individual**

---

[84] *Tesco Stores v Secretary of State for the Environment* [1995] 1 WLR 759.
[85] See L Edwards and M Veale, 'Slave to the Algorithm? Why a 'Right to an Explanation' is Probably Not the Remedy You are Looking For' (2016) *Duke Law & Technology Review* 18.
[86] *Tesco Stores v Secretary of State for the Environment* [1995] 1 WLR 759.

level. To what extent, therefore, should accuracy on aggregate across the system be regarded as relevant in the determination of individual cases? We know, for example, that those who smoke are more likely to develop lung cancer, and that those born in September are more likely to be academically successful, but even if these factors were highly predictive, so that again the ADM can outperform a human, would it be acceptable to use such factors as relevant considerations in making decisions about, for example, entitlement to treatment or A level grades? It is, after all, perfectly possible for non-smokers to develop lung cancer, smokers to escape it, those with August birthdays to achieve at the highest academic levels and those born in September to achieve low grades in exams.

This may, of course, be an instance of the kind noted above where the ADM system is not so much operating differently from a human, as that the way in which it is operating is known and thus more likely to attract review. It is entirely possible that the factors taken into account by human decision makers could be considered relevant as a result of their predictiveness of a particular outcome. Thus, for example, decisions on the licencing of particular drugs or the administration of particular vaccines, will, even if taken by a human, be based on the predicted aggregate impact of that treatment across a population. It may even be that in a decision about an individual factors may be taken into account which are based on statistical predictions such as the likely effect of a drug on the claimant.[87] But it will be necessary for the courts to take a clear view on when such statistical predictions are relevant and when they are not. There may be no alternative, in the making of large cross-population policy decisions such as those relating to drugs and vaccines, to the use of aggregate data, and it may be that while such data is the best available there is no alternative to using it even in an individual case. But it should also be noted that cases both in Australia and the UK have revealed the shortcomings of applying such aggregates in individual contexts,[88] and in such cases the courts should be ready to find the aggregate statistics to be irrelevant.

In addition, however, there are ways in which such systems really will differ from the ways in which a human decision maker would operate. These differences arise from the focus of ADM on correlation as opposed to causation. If a factor is highly correlative with a particular outcome, is that sufficient for that factor to be relevant for the purpose of the ADM decision? For example, a marketing company was able to use YouGov profile data to establish the top ten brands favoured by 'Brexiters' and 'Remainers' in the 2016 Brexit Referendum[89] It is difficult to imagine that the side one took in this referendum could possibly be a

---

[87] See, e.g. *R(Rogers) v Swindon NHS Primary Care Trust* [2006] EWCA Civ 392 where the relevant consideration was the clinician's view that C would benefit from the treatment.

[88] See the problems associated with the Australian Online Compliance Initiative's 'robo-debt' system https://auspublaw.org/2018/04/robo-debt-illegality/ and the decision in *Secretary of State for Work and Pensions v Johnson,* [2020] EWCA Civ 778, albeit on a different legal issue in the latter case.

[89] E James, 'The top 10 brands favoured by Remainers and Brexiters' *Campaign* (1 August 2016) accessed 19 May 2021.

relevant consideration for a public law decision, but let us assume for a moment that it could. If an ADM system were to be created to make this hypothetical decision and it were to make the same calculation that HP sauce was the number one brand for Brexiters and thus purchase of this brand was highly correlated with Brexit voting, would it then be permissible for that ADM system to use HP sauce purchase as a relevant consideration in and of itself? From a human perspective, where we hope that the relevance of our considerations is based on their causal contribution to the desired outcome, such an approach would seem impermissible,[90] and yet if we were to focus only on the accuracy of the system, an ADM system might well outperform a human decision maker as it can do in other contexts. **There is no question that the ground of relevance has the capacity to address and guide the use of ADM systems on this front, but it will need a great deal of more detailed, technical and legal development before it is able to do so.**

4.2.2.5 Issue 5 - Challenging the decision itself - rationality/*Wednesbury* and proportionality

Finally, even if the way in which the ADM system has reached a decision is thought acceptable, there can still be challenges to the decision itself using the grounds of rationality or, as appropriate, proportionality. Although it is generally assumed that proportionality review is more intensive than rationality review, this is in fact not the case[91] and in particular the 'suitability' aspect of the proportionality enquiry, establishing whether the means used were connected to the ends sought, is relatively weak. As long as there is 'not nothing' to connect the means and ends the measure may survive challenge.[92] However, in the rationality context, suitability can in fact be stronger, so that in *Wandsworth*[93] the court struck down a measure for being 'not rationally capable' of achieving the relevant aim and because the connection between means and ends was 'highly speculative'. Similarly, in the *Law Society*[94] case, the measure failed because it was found to do more harm than good. It does therefore seem that such grounds might, for example, provide a helpful way to challenge pernicious feedback loops where the system in fact creates the very thing it is supposed to be detecting.[95] It is also worth noting that if an ADM is discriminatory it will also in all probability not be necessary for proportionality purposes and will fail at that stage of the enquiry. If it is not necessary to treat one group in a disadvantageous way, how can it be necessary for another group?[96] It is helpful that this argument would be applicable in relation to any groups, not just those with characteristics protected

---

[90] Indeed, the inapplicability of the supposed causal line in an individual case was given above as a reason for rejecting the relevance of that consideration in that case, e.g. September birthdays and lung cancer.

[91] See further R Williams 'Structuring Substantive Review' [2017] Public Law 99.

[92] *R (Quila) v Home Sec* [2011] UKSC 45, [2012] 1 AC 621.

[93] *R (Wandsworth) v Schools Adjudicator* [2003] EWHC 2969, [2004] ELR 274.

[94] *R (Law Society) v Legal Services Commission* [2010] EWHC 2550, [2011] ACD 16.

[95] For an example from the private law context, see *Filcams CGIL Bologna, Nidil CGIL Bologna, FILT CGIL Bologna c Deliveroo Italia SRL*, Tribunale Ordinario di Bologna, N RG 2949/2019.

[96] See, e.g. *A v Home Secretary* [2004] UKHL 56, [2005] 2 AC 68.

by legislation. **It is therefore clear again that the grounds of *Wednesbury* unreasonableness/rationality review and proportionality both contain lines of argument that could usefully be used to control and optimise the use of ADM systems by public authorities, but as before, only if this potential is noticed and developed in a suitably interdisciplinary manner by the courts.**

4.3 Beyond public law?

It is clear, then, that there is a great deal of work to be done in developing the grounds of review to be able to respond to the challenges of optimising public authority use of ADM. However, it is also evident that there are grounds which can be developed in this way, and in fact, if we are able to undertake this work effectively, it may even be the case that such grounds could be used more widely than just in context of review of public authorities.

It is well known that there are two touchstones for an entity being considered public for the purposes of judicial review (including review for compliance with human rights); some kind of extraordinary power and the carrying out of a 'public function'.[97] However, as Lord Neuberger pointed out in *YL*,[98]

> The centrally relevant words, 'functions of a public nature', are so imprecise in their meaning that one searches for a policy as an aid to interpretation. The identification of the policy is almost inevitably governed, at least to some extent, by one's notions of what the policy should be, and the policy so identified is then used to justify one's conclusion.

In other words, the question is ultimately a political one; if one is a proponent of big government, more activities will seem like 'functions of a public nature', whereas if one subscribes to small government ideals, then very few things will be 'functions of a public nature'.

If this factor is thus unable to give us an objective test for when an authority ought to be regarded as public, we are left with the other one, the idea of an imbalance of power between the parties as a result of an inherently greater power on the part of the public authority which must, as a result, also carry greater constraints.[99] And of course it is this imbalance of power that is at the heart of the different treatment of public authorities in the GDPR, in all probability preventing them from relying on consent as a ground of lawful processing.[100] But of course public law is not the only context in which such an imbalance of power can occur. It is unsurprising, therefore, that public law concepts such as 'reasonableness' might 'leak' into contexts such as employment law where a similar imbalance occurs,[101] but the question raised by the use of ADM is whether this 'leakage' could occur even more widely. **After all, public law is an**

---

[97] See, e.g. *R v Panel on Takeovers and Mergers, ex p. Datafin* [1987] QB 815.
[98] *YL v Birmingham* [2007] UKHL 27, [2007] 3 W.L.R. 112.
[99] See, e.g. *R v Somerset County Council ex Parte Fewings* [1995] 1 All ER 513 at 524.
[100] Recital 43, discussed above.
[101] See, e.g. *Braganza* [2015] UKSC 17.

area of law specifically designed for the control and optimisation of decision making processes, and so while it does have work to do in developing its toolkit for this new, digital context, if it is able to do that, there is no reason why this toolkit could not be more widely available in other circumstances where the ability to use big data gives even a private entity a similar kind of extraordinary power to that wielded by public authorities. It is also worth noting that in its regulation of even private entities the ICO is subject to judicial review,[102] giving a second possible channel through which public law might have the potential to influence even private contexts.[103]

4.4. Procurement

As discussed above, the majority of public sector uses of ADM are likely to involve at least some procurement of services. This could mean procuring an entire stand-alone ADM service from a single provider, or the outsourcing of multiple elements, such as cloud computing services for building and deploying algorithms, from multiple providers. Since many of the key technical, legal and governance factors of an ADM are effectively determined and constrained by decisions made during the procurement process, this is a key area to consider.

Several efforts to shape and promote the use of ADM within government have focused on the procurement process. The World Economic Forum, in partnership with the UK Office for AI, have produced guidelines on AI procurement. These consist of high-level recommendations, relating to various ethical principles and governance considerations, as well as emphasising the need to 'incorporate legislation and codes of practice' into procurement processes.[104] While public authorities can procure ADM services independently, they may also do so through the Crown Commercial Service, the Cabinet Office trading fund which leads procurement policy on behalf of government. The CCS has set up a Digital Marketplace which is responsible for around 25% of public sector technology procurement.[105] Last year, in collaboration with the Office for AI, CCS released a

---

[102] Those concerned by a particular ADM decision will refer this issue to the ICO. The initial right of recourse against the ICO is to the First Tier Tribunal (General Regulatory Chamber), but DPA 2018 ~~S~~s 163(3) suggests that the grounds of judicial review will be relevant as potential grounds for allowing an appeal. Appeal then goes up through the Upper Tribunal,~~-~~ ultimately, though rarely, leading to judicial review. See~~;~~ e.g. ICO Disclosure Log – Response IRQ0686975.

[103] See further R Williams, 'Rethinking Administrative Law for Automated Decision Making', *Oxford Journal of Legal Studies*, forthcoming.

[104] https://www.weforum.org/reports/ai-procurement-in-a-box/read-the-reports. The guidelines contain 10 high-level recommendations: 1) Focus on outlining the problems and opportunities they're looking to solve, and use procurement vehicles that promote iterative solution development. 2) Define the public benefit of using AI, while assessing risks. 3) Align with relevant existing government strategies and contribute to their improvement. 4) Incorporate relevant legislation and codes of practice. 5) Articulate the feasibility of accessing data that may be relevant for the AI solution. 6) Highlight any limitations of potential data. 7) Work with a diverse, multidisciplinary team. 8) Focus on algorithmic accountability and transparency. 9)Plan for the support that will be required to operate the tool throughout its life cycle. 10) Create a fair, level playing field among AI-solution providers.

[105] Artificial Intelligence and Public Standards Report

new ‘Artificial Intelligence Dynamic Purchasing System (DPS)’ which aims to make it ‘easier to buy [AI] and take advantage of what it has to offer’.

The role of such procurement mechanisms in setting standards around public sector use of ADM is emphasised in a 2019 report from the Committee on Standards in Public Life on the impact of AI on public standards, which argued that the public procurement marketplace should integrate ethical standards and require vendors to evidence how they meet them.[106] This concern was prompted by private AI providers who had the capability to make the ADM systems explainable, “but that they were rarely asked to do so by those procuring technology for the public sector ... The Committee was told that requirements for technical transparency are not usually included in procurement tenders and contracts."

The DPS goes some way towards remedying this situation, by bringing the GDS Data Ethics framework and Office for AI’s AI procurement standards into the ordering process, and ‘addresses ethical considerations when innovating and buying artificial intelligence and was designed to help customers build in a strong ethics process’.[107] Suppliers on the DPS are required to describe their approach to “data limitations, fairness and bias”, “diversity of teams”, data protection, the “level of human decision-making at critical points”, and how they will make their systems “transparent and explainable to a non-expert audience”.

While the principles of these procurement frameworks do reflect some parts of the relevant legal frameworks applicable to ADM systems, including data protection and equality,[108] they fall short of a full consideration of the legal obligations on public bodies, particularly those stemming from administrative law. Some of these more detailed legal considerations may be covered in the broader framework of procurement rules and regulations, while these recent AI-specific procurement guidelines aim only to emphasise key considerations. As the Office for AI’s AI Procurement in a Box toolkit states, it is assumed that “the reader has a sound working knowledge of those rules and of the end-to-end procurement process". Where legal compliance is raised, it is as a reminder rather than at the core of the framework; e.g. “remember to make sure your use of automated decision-making does not conflict with any other laws or regulations”.[109]

The ADM-specific implications of these more detailed legal considerations may be partly dealt with in procurement framework schedules which are used when suppliers sign contracts with CCS. For instance, Joint Schedule 11 provides data

---

[106] Ibid
[107] Dynamic Purchasing System Marketplace
[108] For instance, the OAI Guidance on AI Procurement advises public services procuring AI systems to give “due considerations required by the Public Services (Social Value) Act 2012 (as amended), as applicable, and assessing application of the Public Sector Equality Duty under the Equality Act 2010”
[109] A guide to using artificial intelligence in the public sector

protection related definitions such as controller / processor.[110] These can be incorporated and modified to suit the particular context of the project. However, the broader procurement framework is arguably more focused on ensuring competitive tendering and value for public money, than ensuring public law values are considered at the earliest stage of AI procurement. **Moreover, there is a heavy focus on *ethics*, while requirements of public law appear to be relatively under-emphasised.** Despite the challenges algorithms pose to compliance with the standards of public law, the focus is on ethical standards.

Another potential issue with using a Dynamic Purchasing System for ADM procurement is that it may be better suited to purchasing of simple commodity services, which can be evaluated along simple standardised criteria. The use of ADM in the public sector raises a huge range of difficult and nuanced legal questions; seemingly small differences in the design of an ADM system could have significant implications for legal compliance. **A DPS may not allow for the kind of nuanced, case-by-case assessment and comparison needed to apply public law principles within the procurement process.**

5. Regulatory Models

Previous sections suggest that existing legal regimes, including DP, equality law and judicial review, may have significant gaps or areas for development. In this section we review alternative regulatory models which have been proposed elsewhere to address challenges of ADM, which might provide inspiration for future legislative approaches to ADM in the public sector.

5.1 The proposed EU AI Regulation

On April 21 2021, the EU Commission adopted a proposal for a regulation (the "AI Regulation") on "artificial intelligence systems" (AI systems), which it describes as "the first ever legal framework on AI[111]." We do not plan to give a full summary of the AI Regulation but to point out some aspects of its approach which may be useful to think about in the broader context of developing legislative oversight for public sector ADM systems.

Wide definition of AI; public sector relevance of "high risk" AI.
The AI regulation casts a wide net, capturing not only AI systems offered as stand-alone software products, but also products and services relying on AI

---

110 https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/790967/Joint_Schedule_11__Processing_Data__v.4.0.pdf

111 See Proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) Brussels, 21.4.2021 COM(2021) 206 final https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence-artificial-intelligence .

services directly or indirectly. “AI systems” extend by Annex 2 to software which employs any of the following techniques or approaches:

- Machine learning (including supervised, unsupervised and reinforcement learning, using a wide variety of methods including deep learning).
- Logic- and knowledge-based approaches (including knowledge representation, inductive (logic) programming, knowledge bases, inference/deductive engines, (symbolic) reasoning and expert systems).
- Statistical approaches, Bayesian estimation, search and optimization methods.

The AI Reg is in principle much wider than the GDPR because it is not constrained in scope to the processing of personal data. This can remove artificial hurdles to considering both non-personal and personal data aspects of a system. Its scope is also far far wider than the unhelpful limitations in art 22 (and by implication, arts 13-15) relating to “solely automated decisions”.

However the thrust of the AI Regulation is in fact mainly aimed at “high risk” AI systems which are much more closely defined as :

- “safety components” of products, or
- products
    - covered in EU legislation listed in Annex II (eg, machinery, toys, lifts/elevators, radio equipment, pressure equipment, marine equipment, cableways, gas-burning appliances, and medical devices), and
    - AI systems listed in Annex III (related to biometric identification and categorization of natural persons; management and operation of critical infrastructure; education and vocational training; employment, works management and access to self-employment; access to and enjoyment of essential private services and public services and benefits; law enforcement; migration, asylum and border control management; and administration of justice and democratic processes)

As can be seen, many “high risk” AI systems are likely to be partly or wholly operated by the public sector.

Obligations for all “high risk” AI systems

- Data and data governance (see below).
- Technical documentation and record-keeping.
- Transparency and provision of information.
- Human oversight.
- Accuracy, robustness, and cybersecurity.

There is much conceptual overlap with the GDPR here, remembering of course that the AI Reg is not restricted as the GDPR is to systems which process personal data. But some aspects of the AI Reg are novel and may “fix” some of the issues discussed in earlier sections.

Register of high risk AI systems to be maintained.
This should contain information on the provider, the AI system trade name and any additional unambiguous reference allowing identification and traceability of the AI system as well as a description of the intended purpose of the AI system. **We have noticed already that neither DP nor public law mandate such a register for public sector ADM systems despite frequent calls for such by politicians, advocacy groups and adoption by some European city councils.**[112]

Conformity certification.
Prior scrutiny and post hoc accountability of AI systems are extremely labour intensive for individual public authorities to guarantee on an ad hoc basis. We have already noted the usefulness of the DPIA and the Equality Impact Assessment in making system operators think in advance about the impact of their system on user rights and freedoms as well as about how to mitigate these risks. However we have also highlighted how systems increasingly draw on components including training sets and partly or wholly trained models supplied by various private sector vendors. Maintaining oversight and control over these different moving parts is a very tough call for the resource-strapped public sector.

Montoya and Rummery note that "the blurring of the public and private has implications for the amount of control government is able to exercise over the design and specification of [ADM systems]" and that "having outsourced the design and coding of a system to the private sector, government may not possess the technical capability to assess any potential impact of the system before or after its deployment". They warn that establishing robust accountability might "by necessity be "resource- and time-intensive, a requirement at odds with the speed and scale at which these systems are being adopted".

The approach of the AI regulation to this problem is loosely modelled on the EU's existing Product Safety regime and is to systematically require every actor involved in the development, distribution and use (or re-use) of relevant AI systems to meet certain levels of conformity with detailed regulation. **Conformity is either self assessed or for a limited set of very high stakes systems, externally audited, and is kitemarked by a "CE" mark. Thus, *all providers of high risk AI systems* are responsible for ensuring the compliance of their systems with the AI Regulation.** Manufacturers of products including high-risk AI systems, are also responsible for compliance as if they were the provider of the high-risk AI system. Distributors, importers, users and other third parties are also subject to some extent to the same obligations as providers.

---

[112] E.g. the Institute for the Future of Work (https://www.ifow.org/publications/mind-the-gap-the-final-report-of-the-equality-task-force), the Ada Lovelace Institute (https://www.adalovelaceinstitute.org/event/mandatory-reporting-public-sector-algorithmic-accountability/), Dutch MP Kees Verhoeven (https://algorithmwatch.org/en/kees-verhoeven-algorithm-registry/). Such registers have been implemented at a city level in Amsterdam, Helsinki, and Nantes (https://www.adalovelaceinstitute.org/event/mandatory-reporting-public-sector-algorithmic-accountability/).

The net effect is a general ability to identify an AI system, or component, which meets quality criteria by its CE mark, just as currently we can identify a safe product (eg, an imported toy) by its EU CE kitemark. Although there are many issues with this scheme, the broad brush is an attempt to create a systematic web of trust and compliance in AI products which the public sector could then rely on when procuring or building systems in a global market of public/private partnership.

Training set data

One of the regulatory requirements for the CE mark is that training, validation and testing datasets must meet certain quality criteria (draft art 10). This is a crucial aspect of building a fair, inclusive, accurate and unbiased system so far largely untouched by specific regulation, as highlighted in section 1. Although the particular criteria can be debated, it is notable that the suggestions in draft article 10 are both wider and more detailed than those usually mentioned in connection with DP law and human rights, and include practical matters of design and engineering, and forefront complete representation of groups rather than the individual interests of the GDPR.

Human oversight

Draft art 14 provides that high-risk AI systems shall be designed and developed in such a way that they can be "effectively overseen by natural persons" while in use. The aim is to mitigate issues of opacity and "black box" systems so that humans charged with oversight of systems can:

- Understand the system sufficient to monitor for anomalies and dysfunctions
- Guard against "automation bias"
- Be able to correctly interpret system output and
- Be empowered to know when and how to disregard or override the system.

Much of this may seem aspirational to say the least and there is a worrying dependence in this article on the original provider of the AI system identifying how to meet these criteria, when it is quite likely that the system performance will change entirely when put into use by a buyer or operator and when it starts to learn. However again, it is the polar opposite to the very limited controls and transparency imposed on a very small subset of systems by art 22 GDPR.

Banned systems

The AI Reg is the first major piece of proposed EU legislation to consider that certain AI systems should simply not be developed. We noted above that certain systems simply cannot act accurately, to, eg, predict major life outcomes over time and at least at present would be both unethical and wasteful to build. Other systems are coming to be regarded as morally unacceptable or at best as highly prone to abuse. Face recognition in law enforcement hands is one such contentious example and above we noted the *Bridges* case as the major authority we so far have on the legality of public sector ADM. This has been so far the most controversial part of the AI Reg and only four examples are

included: subliminal manipulative systems; systems which exploit vulnerabilities related to age, and physical or mental disability to distort behavior; public sector "social credit" systems; and real time remote biometric systems in public spaces. All of these are hedged with exceptions and criteria (eg, causing physical or psychological harm) and the first two are hard to connect to systems actually in operation; but they are mentioned here as an example of a precautionary approach which new laws could take[113].

5.2 The Digital Services Act

While aimed at regulating online platforms rather than public sector algorithms, the proposed EU Digital Services Act 'package' (consisting of the consumer-oriented Digital Services Act and the competition-oriented Digital Markets Act) provide examples of regulatory models which could be adapted to the contexts discussed above.

Under the Digital Services Act, 'very large online platforms' (VLOPs) are subject to two duties which would enable greater scrutiny of the use of algorithmic decision-making. First, under Article 28, VLOPs must pay for independent audits of their systems, to ensure compliance with protocols and codes of conduct developed as required under various other substantive obligations of the Act. The Article assumes the development of a class of specialised auditors to undertake such work, and sets out conditions to ensure their independence from the VLOPs obliged to commission them. This independent audit regime is complemented by Article 31, which requires VLOPs to give regulators access to data necessary to monitor compliance with the Act. This data and the technical processes for scrutinising it will be further specified through delegated acts adopted by the Commission.

5.3 Impact Assessments and *ex ante* Accountability

Several legislative proposals and initiatives in various jurisdictions look to ex ante accountability mechanisms such as impact assessments[114]. These can be seen as consonant with the DPIA regime discussed above in section 3.6.1, but may differ in their scope and role within a broader enforcement regime. Broadly, impact assessment requirements aim to identify possible risks arising from the use of ADM, quantify the likelihood and severity of the harms associated with them, and document the development and application of measures to mitigate those risks to bring them within a tolerable risk threshold. In some cases, these involve prescribed types of risk and mitigations to consider. For instance, the Canadian Treasury Board released a mandatory algorithm impact assessment tool, to support the Treasury's Directive on Automated Decision-Making, which

[113] For a comprehensive critique of the draft EU Regulation;s idea of "impermissible uses" see EDRi *Recommendations for a Fundamental Rights-based AI Regulation*, June 2020 at https://edri.org/wp-content/uploads/2020/06/AI_EDRiRecommendations.pdf .
[114] See further work of Ada Lovelace Institute, https://www.adalovelaceinstitute.org/project/algorithmic-accountability-public-sector/.

specifies 48 possible risks and 33 types of mitigation.[115] This is in contrast to the likes of the GDPR's impact assessment requirement, which does not specify risk types, but rather requires anticipation of any risks 'to the rights and freedoms of natural persons'. As discussed above, the public sector equality duty (PSED) requires public authorities to have due regard to the need to eliminate unlawful discrimination; equality impact assessments (EIAs) are a common way in which they document how they have discharged this duty.

Based on an analysis of shortcomings in data protection and equality law, a recent report from the Institute for the Future of Work's Equality Taskforce has argued for the need for a new set of ex ante duties.[116] These build on an earlier framework for equality impact assessment for algorithms, and include: new statutory duties for public consultation, reasonable adjustments, and co-operation across the ADM supply chain; public disclosure of impact assessments and ongoing data on outcomes; collective rights to know and be involved (e.g. for trade unions); a series of clarifications to existing equality and data protection law to clarify their application to ADM; and greater coordination between regulators including jointly-issued statutory guidance. While principally focused on equality in the work context, this proposal provides additional forms of ex ante governance which potentially have broader applicability to the public sector.

5.4 Summary

Several proposals for novel ADM legislation suggest the potential for alternative regulatory models. While different, they all have in common the aim of pushing the governance of ADM systems *further up* the supply chain and *earlier on* in the lifecycle of the development and deployment of ADM systems. Whether this is through certification of ADM systems prior to deployment as envisioned under the proposed EU AI Regulation, or through ex ante consultation and impact assessment, these forward-looking controls and obligations provide a complement to the largely backward-looking approach of judicial review. **These measures could ultimately strengthen the current regimes by setting out clearer processes through which public authorities can justify their use of ADM (in data protection, equality and public law), and clarify the appropriate processes by which claimants can challenge such uses;** rather than cases being mired in uncertainty arising from the status quo, which is arguably a complex, gap-filled patchwork.

115

https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html

116

https://www.ifow.org/publications/mind-the-gap-the-final-report-of-the-equality-task-force

Appendix A: Issues, Limitations and Risks of ADM Systems

This report has focused on explaining the existing legal and regulatory frameworks governing the use of Automated Decision Making and Assisted Decision Making by public sector bodies. This section provides an overview of some of the issues and risks which should be discussed and mitigated for when developing and deploying ADM systems.

Error

Perhaps the most important and straightforward is that ADM systems never work perfectly, and often do not work as well as claimed by their vendors.

To some extent, errors are inevitable. For rule-based systems, there will always be exceptions to rules which only become apparent when we are confronted with particular cases and extenuating circumstances; exceptions which human case workers may have the capacity to recognise but which an algorithm needs to be explicitly programmed to handle. Frequently, problems for rule-based systems arise when input variables are measured in ways which may sound reasonable, but fail to account for exceptions arising in practice. Errors in the Universal Credit system due to the way monthly income is calculated, which failed to account for differences between HMRC salary data and DWP calendars, are one high-profile example.[117]

For statistical systems, errors are also endemic, but for a different reason. Their purpose is to draw generalisations from data which are mostly true. In most cases, there will be some examples in the training data which go against a general trend, but statistical modelling and machine learning algorithms aim to capture the general trends, not the exceptions.

A related concern is due to the way that statistical models are typically designed to be the best ‘fit’ for the data they have been trained on. This presents problems where the population is heterogeneous; if two groups differ in size and in their distributions of features, a statistical model will typically sacrifice accuracy on the smaller group in favor of better performance on the larger group, to achieve higher overall accuracy. As a result, the statistical ‘majority’ population are more likely to receive the correct result. This may be partly addressed by collecting more data - either more features or more samples which allow the model to perform better on the statistical ‘minority’ - but is an endemic feature of statistical modelling.

117

https://www.hrw.org/report/2020/09/29/automated-hardship/how-tech-driven-overhaul-uks-social-security-system-worsens#_ftn138

There are also important concerns raised by the balance of different *kinds* of errors. Someone being incorrectly denied a benefit as a result of a fraud risk detection model (a false positive error), has quite different consequences, both for the individual and for DWP, to someone being incorrectly granted a benefit (a false negative error). Statistical models always face a trade-off between different kinds of error, and this trade-off is explicitly addressed in the modelling process. Managing these different kinds of error is like squeezing a balloon - avoiding one type of error always comes at the cost of increasing another, although not necessarily by an equivalent degree. This is the same dilemma faced by the justice system, where the calibration between false positives and negatives is the subject of constant debate; as explicitly acknowledged in Blackstone's maxim, that it is 'better that ten guilty persons escape, than that one innocent suffer'. Data scientists can effectively set what these exact proportions ought to be when they design a model, and so long as the data they test the system is representative of the deployment context, can do so with relative confidence that the balance of errors will be distributed accordingly.

The ability to accurately measure expected errors will differ depending on the various approaches to integration as discussed above. In a scenario where the ADM system is developed in-house, using training data drawn from existing services and operations, the system can be tested on data from the same source. This means that an in-house digital team can be more confident that the level of accuracy reported in testing will match accuracy in live deployment. However, when part or all of the ADM development process is outsourced, it becomes harder for the customer to assess how accurate the system will be when deployed in their context. ADM providers might not report the results of their own testing, or only report accuracy in aggregate and not provide a breakdown of different types of error. Moreover, the results of any testing that the ADM provider undertakes may not apply when deployed in the customer's context, if the data used to train and test the ADM system differs statistically from deployment context. For instance, the reported accuracy of a facial recognition system trained on a given population, and tested under a given set of conditions (e.g. passport gates), may not hold when the system is deployed on a different population under different conditions (e.g. CCTV footage). As such, customers of outsourced ADM systems cannot take a vendor's reported accuracy at face value and would need to conduct independent testing to have confidence in how accurate the system is in their context.

Discrimination and equality

There are several reasons why an ADM system might unjustly discriminate between groups with different characteristics (including the protected characteristics of equality law, but potentially others which are not (yet) included, such as socio-economic status).

One reason is already discussed above; because statistical models typically seek to minimise overall error, they are most accurate on statistical majorities. In

so far as the statistical majority corresponds to a protected characteristic (e.g. a majority white population), a model trained even on a ‘representative’ data set will perform worse on people without that protected characteristic.

Another reason is that the choice to use particular kinds of data in prediction, or the measurement of the outcome being predicted, might systematically under or over-estimate the true value for certain protected groups (often called ‘measurement bias’). For instance, consider a model which uses an individual’s number of prior drug offences to predict future re-offence. The measurement of the predictor variable (prior offences) and the target variable (future offences) may be systematically biased as a result of racially unequal stop and search. If black people are more likely to be stopped and searched than white people, such a system will over-predict re-offence among the black population. Even if the re-offending rates are the same between groups, and even without explicitly factoring in race, such a system would indirectly discriminate against black people. Or, using an employee’s sales figures as a proxy for their productivity might misrepresent the productivity of women employees, if they tend to undertake a larger share of productive work tasks which don’t directly result in sales.

A further cause of discrimination in ADM systems is underlying differences in the distribution of features in the population. For instance, there may be differences in rates of reoffending, school exclusion, or fraud. The reasons for such differences are typically complex and varied, subject to debate by social scientists, and cannot be inferred from the data alone (and certainly not by an algorithm). Often, they are the result of systemic social inequalities between groups. Even if a statistical model does not directly use protected characteristics as an input, any group-level inequalities are very likely to be reflected in the predictions the model makes, and therefore have unequal effects between protected characteristics. But a higher prevalence of some outcome among those with a protected characteristic doesn’t justify the use of an ADM system which treats people with that characteristic less favorably; doing so arguably unfairly penalises individuals for a statistical feature of the data which they have no control over.

Various methods exist to assess and mitigate forms of indirect discrimination that can arise from statistical models. Assessment typically involves comparing outputs and errors of the ADM system when applied to a set of test data, where the protected characteristics of individuals in the test data are known, to identify significant differences between protected groups. Mitigation may involve changing the training data or modelling process to reduce the influence of protected characteristics on the outputs of the model before deployment, or potentially using an individual’s protected characteristic as an input to the decision itself during deployment. In either case, data about protected characteristics (either of the individuals in the test data, or of decision subjects during deployment) would be required.

Robustness, generalisation and feedback loops

A fundamental tension exists in statistical modelling between models performing well on data drawn from one context, and performance on data from different contexts. For instance, a speech-to-text model might be designed to perform better on accents which match the accent(s) of those it was trained on, or to perform less well overall but better on a wider variety of accents, including those it hasn't 'heard' before. The same tradeoff affects all kinds of statistical modelling, from risk assessments to document classification; they can perform better in contexts that match the training data but worse on those which don't, or they can perform less well in the 'known' contexts but better overall on the 'unknown' contexts. This also applies over time; as the world changes, the context from which the model's training data were drawn changes, so a model designed to best fit the training data may fail faster than one designed to generalise to many contexts (this is known as 'overfitting' or variance error). The opposite problem, where the model fails to capture enough of the patterns in the context of the training data, is known as 'underfitting' or 'bias error'.

The ability of an ADM to work well in different contexts, where the distribution of features and outcomes differ to those in the training data, is known as generalisability, and sometimes as 'robustness' (although robustness has other meanings, discussed below). Several factors make it challenging to ensure that the ADM systems in use in the public sector are generalisable, especially where ADM systems are not developed in house. This is because generalisation is best measured by testing the model on data from different distributions, and best achieved by including such data in the model's training. But in outsourcing relationships, the system vendor may not have access to data from the context of deployment, and the customer may not have the ability (technically or contractually) to test and re-train the model accordingly. Models and datasets developed for private sector uses might be plausibly expected to generalise to public sector uses (for instance, a risk scoring model used to evaluate bank accounts for anti-money-laundering purposes might be repurposed for the purposes of detecting welfare benefit fraud), but establishing this might require levels of coordination and integration between the public authority and the private provider which raise problems of their own, including concerns about data protection and privatisation of services.

Robustness also refers to a general property of algorithmic systems as performing within expected boundaries of behaviour even in the face of unusual inputs or unexpected contexts. This could be due to unusual or erroneous input variables resulting from the failure of another piece of software or data source, or as a result of a deliberate attempt to subvert or attack the system. While the latter may be (as yet) uncommon in the public sector algorithm context, the former is likely to be a problem, especially where algorithms are intended to be plugged into the (necessarily) complex back-end systems and bureaucratic processes of central and local governments.

Feedback loops

Related to robustness is the problem that statistical systems which are intended not only to observe and predict, but also intervene on the world, often end up undermining their own operation. This is because if the systems' predictions and classifications are used to influence decisions, then a dynamic system is likely to respond to those decisions in ways which mean the training data may no longer capture the new patterns of behaviour. This can lead to issues such as feedback loops, where by following the algorithms' predictions, decision-makers create more evidence for certain types of outcomes and miss evidence for other outcomes. For instance, if predictive policing systems dictate where officers go, those areas may inevitably end up with higher recorded rates of crime simply because there are more officers around to arrest people; for this reason, predictive policing has been said to predict *policing*, rather than crime.[118]

Limits to prediction

While algorithms have the potential to help in many areas of the public sector, if designed and tested appropriately, there are some problems which are inherently unpredictable and for which statistical models cannot reasonably be expected to help. While recent years have shown that a surprising range of prediction problems can be solved with lots of data and machine learning algorithms, many commercial applications of AI simply cannot predict what they claim to be able to predict with sufficient accuracy.

This is particularly true when it comes to predicting major life outcomes of individuals over time, such as a child's performance at school, whether a household will be evicted, or whether someone will be made unemployed. A study published in 2020 and led by researchers at Princeton recruited 160 teams of data scientists and social scientists, and gave each team the same set of detailed, longitudinal data about families over a 15 year period.[119] Despite the high quality data, expertise and resources available to the teams, none were able to create models which could predict life trajectories with a high level of accuracy. These are prime examples of the kinds of outcomes that public authorities might hope to be able to predict using statistical models, and there are strong incentives for commercial providers to attempt to provide such models. However, the results of this research suggest that such outcomes may be simply beyond the limits of prediction.

Individual level accuracy

The accuracy of statistical models is typically only measurable at a population or group level. For instance, people classified as 'high risk' of becoming unemployed by a risk scoring model might actually become unemployed 20% of the time, vs 2% of people in the 'low risk' category. But while it is possible to

[118] Lum, Kristian, and William Isaac. "To predict and serve?." *Significance* 13.5 (2016): 14-19.
[119] Salganik, Matthew J., et al. "Measuring the predictability of life outcomes with a scientific mass collaboration." *Proceedings of the National Academy of Sciences* 117.15 (2020): 8398-8403.

measure the chance of becoming unemployed within each category, that doesn't account for variation between individuals within each category. An individual's chance of becoming unemployed is not the same kind of probability as that involved in their chances of winning the lottery. Even if 20% of the high-risk category can reliably be predicted to become unemployed, are they all equally likely to have that outcome or is there something about the 20% that make them more likely?

It is typically impossible to assess individual-level accuracy from the data alone, unless multiple repeated predictions are made of an individual, and their subsequent outcomes are observed. In the rare cases where individual-level accuracy has been measured, such as with the use of risk assessments in bail hearings in the US, it is possible to observe the difference between individual and population-level accuracy.[120] This shows how a system could be highly accurate, and reflect the real distribution of risk at a population level (in statistical terms, it is 'well calibrated'), and yet still fail to reflect the true risk at an individual level. In other words, individuals classed as 'high risk' may not differ greatly from those classed as 'low risk', even if the model performs well at a population level. While statisticians often warn decision-makers using their systems that population-level risk is not the same as individual risk, the predictions that models are often treated as if they are accurate at the individual level.

Uncertainties about individual level accuracy of a statistical model were one of the issues arising in relation to the Ofqual algorithm originally proposed to award A-level grades to students in England and Wales in summer 2020. In the initial approach, which was later reversed, student grades were determined by three inputs: first, teachers' assessments of what grades the student would have got in each subject; second, the teacher's assessment of each student's relative rank within their year group for that subject; a statistical model which decided how many grades at each level would be distributed to each school, based on prior performance. For large cohorts, an algorithm then assigned the allocated grades per school to students in rank order. For small cohorts, teacher's assessments were used (since teacher's assessments were more generous, this favored smaller cohorts). The assumption was that teachers would be able to accurately rank students in relation to each other, but might over-estimate the grades they would achieve. However, even if the model had been highly accurate in its predictions about the distribution of grades between schools, it is unclear how accurate the distribution of grades would have been *within* schools - i.e., at the level of individual students. The only way to estimate this would have been by comparing teachers' expected rankings against actual rankings from previous years, but such data did not exist.

Similarly, the use of Covid risk assessment tools, which place individuals into risk categories on the basis of various demographic and self-assessed physiological

[120] Lum, Kristian, David B. Dunson, and James Johndrow. "Closer than they appear: A Bayesian perspective on individual-level heterogeneity in risk assessment." *arXiv preprint arXiv:2102.01135* (2021).

data, might be accurate on a group-level, but could hide large within-group variance.[121]

Unobserved labels

Often, statistical models are deployed where a decision-maker wants to be able to predict an outcome under different courses of action, for the purpose of choosing between those actions. For instance, what will happen if this kind of person is granted a loan, or what would we find if we investigated this taxpayer for fraud?

However, since models are trained on historical data, if a given course of action has never been taken, they have no basis on which to predict the outcome for such courses of action. From the perspective of a data scientist designing a statistical model, the training data would ideally include examples of randomly allocated interventions, i.e. where loans are granted at random regardless of credit risk, or citizens are investigated for welfare fraud at random. This would ensure that there are no 'gaps' in the dataset, where we lack ground truth labels for certain types of cases. Of course, in high-stakes decisions, there are good reasons why this is not possible, for instance, banks are unlikely to grant loans to low credit score applicants. However, this means that statistical models are typically informed by only a subset of relevant possibilities; unlike the kind of randomised control trials used in medical research, where effects are randomly distributed among the trial population.

As a result, statistical models have large blindspots where the data that would allow them to predict certain outcomes simply doesn't exist. Even if past loan decisions, or tax fraud investigations have been to some extent well-targeted, it is impossible for us to verify this because we can't observe what would have happened given different historic decisions. This means that when it comes to making predictions about individuals who are similar to those who were historically not granted loans, or not investigated for tax fraud, the decision-maker can only infer from the model what will happen under the same pattern of treatment; the results of the alternative treatment cannot be predicted by the model with any confidence. However, this inconvenient limitation is often neglected when models are deployed in real-world contexts, and is not included in claims about how accurate a system is.

Opacity, transparency, explainability

Much has been made of the supposedly 'black box' nature of modern AI systems, where not even their designers understand how they are making decisions. However, this problem really only relates to the technical complexity of a limited range of novel techniques which are typically not deployed in public sector settings. Even where they are deployed in public sector settings, there may only be a small number of cases in which the gains in accuracy are worth

[121] Sundar, Santhanam. "Covid-19 risk assessment: a futile metaphorical strip search." *bmj* 370 (2020).

the additional costs of complexity and resulting opacity. In many cases, simpler, inherently explainable models can be built with very little or no loss of accuracy. While the less inherently interpretable approach of 'deep learning' has proven particularly useful for computer vision problems, it often does not significantly outperform more inherently interpretable models. For instance, work at the intersection of computer science and statistics has shown that inherently interpretable 'scorecard' models can be automatically created for applications such as credit scoring and recidivism risk assessment without paying any penalty in terms of accuracy.[122] Where a complex model really is needed, there may be other methods of explaining how the system works. These can be at the level of an individual decision, for instance, by decomposing the different inputs to a decision and showing how each contributed to the output.

The broader issues of transparency are complex and stem from more than just technical limitations. The supply chains of ADM systems are perhaps more complex and more in need of explanation and transparency than the way that particular outputs are reached. As explained above, the procurement of training data, modelling processes, the incorporation of models into software, and the incorporation of software into organisations, are often undertaken by dispersed teams between and within organisations. The ability to explain how all those elements fit together, and show their provenance, is likely to be hampered by the difficulties of communicating those things across the supply chain, contractual barriers, and intellectual property concerns.

Correlation and causation

Most kinds of statistical modelling, including machine learning, deal with correlation rather than causation. The human mind excels at imagining and hypothesising about the possible causal relationships between variables, but traditional statistical models are incapable of doing this. While they can identify when A is correlated with B, machine learning algorithms are incapable of inferring whether A causes B or vice-versa, or whether there is some third, unmeasured variable, which causes both A and B, or indeed, if the correlation is purely coincidental.

This problem is inherent in the opportunistic approach taken in much of machine learning, to simply use whatever data you have available. This is unlike much of science, where specific data generated in a controlled laboratory setting, or sought out in field studies and natural experiments, which allows causal hypotheses to be tested and normally co-occuring variables to be teased apart.

In this sense, even the simplest of statistical models are a kind of causal black box when it comes to explaining why they make the predictions they do. We can point to the correlations they have observed in the training data, but we can't explain why those correlations hold. It may be that they are causally related, they

[122] Rudin, Cynthia, and Joanna Radin. "Why are we using black box models in AI when we don't need to? A lesson from an explainable AI competition." *Harvard Data Science Review* 1.2 (2019).

might both be caused by some other factor, or it might be just a spurious correlation. We may be able to explain how the model was built from the data, but we don't necessarily know why the data is the way it is - that demands an explanation in causal terms.

Automation bias, rigidity and overdelegation

As explained above, ADM systems can be designed as decision-support systems which form just one factor alongside others in a human decision making process, or as ‘autonomous’ decision-making systems which trigger material effects without human intervention. This distinction is not one that can be settled by stipulation or by design ab initio, but emerges as a result of how a system gets interpreted, embedded and shaped by the people who use it over time. This means that a system initially intended as decision-support may end up effectively automated if the human decision-makers routinely defer to it and eschew their own judgement. Conversely a system that was originally sold as taking decisions automatically may in fact rely on (often hidden) human labour to make up for its initial or endemic deficiencies, or which emerge during deployment (a phenomenon called ‘fauxtomation’).[123]

The term ‘automation bias’ describes the phenomenon where human decision-makers come to defer to the outputs of what was intended to be merely a decision-support tool. This could be caused by automation-induced complacency, where the system is so often correct (or perceived to be correct), that the human decision-maker loses interest, fails to pay attention, due to being insufficiently engaged by the task. Or it might be the result of the human decision-maker having insufficient authority or trust from management to take a decision which runs counter to the ADM, perhaps due to expectations about where blame and liability will lie if decisions turn out to be incorrect. The converse may also occur, where the ADM system is so frequently incorrect, that the decision-maker ceases to pay attention to it all, ignoring useful evidence from the system.

[123] https://logicmag.io/failure/the-automation-charade/

The Legal Education Foundation
Registered office
Suite 2, Ground floor
River House
Shalford, Guildford
Surrey GU4 8EP

www.thelegaleducationfoundation.org

Registered charity no. 271297
Registered in England and Wales

The
Legal
Education
Foundation

Designed and produced by webb-jenkins.com