Administration by Algorithm?

Public Management meets Public Sector Machine Learning

Michael Veale[1, 2, 3] and Irina Brass[1]

[1]Dept of Science, Technology, Engineering & Public Policy, University College London
[2]Birmingham Law School, University of Birmingham
[3]The Alan Turing Institute, London

This is a draft of a chapter that has been accepted for publication by Oxford University Press in the forthcoming book *Algorithmic Regulation* edited by Karen Yeung and Martin Lodge, in press 2019.

Correspondence to: mveale [at] turing.ac.uk.

Public bodies and agencies increasingly seek to use new forms of data analysis in order to provide 'better public services'. These reforms have consisted of digital service transformations generally aimed at 'improving the experience of the citizen', 'making government more efficient' and 'boosting business and the wider economy'. More recently however, there has been a push to use administrative data to build algorithmic models, often using machine learning, to help make day-to-day operational decisions in the management and delivery of public services rather than providing general policy evidence. This chapter asks several questions relating to this. What are the drivers of these new approaches? Is public sector machine learning a smooth continuation of e-Government, or does it pose fundamentally different challenge to practices of public administration? And how are public management decisions and practices at different levels enacted when machine learning solutions are implemented in the public sector? Focussing on different levels of government: the macro, the meso, and the 'street-level', we map out and analyse the current efforts to frame and standardise machine learning in the public sector, noting that they raise several concerns around the skills, capacities, processes and practices governments currently employ. The forms of these are likely to have value-laden, political consequences worthy of significant scholarly attention.

Contents

1 Introduction

Public bodies and agencies increasingly seek to use new forms of data analysis in order to provide 'better public services'. These reforms have consisted of digital service transformations such as 'e-government 2.0' and the creation of 'integrated data infrastructures' (New Zealand 2016), generally aimed at 'improving the experience of the citizen', 'making government more efficient' and 'boosting business and the wider economy' (Manzoni, 2017).

It is not a new observation that administrative data — data collected by or for public bodies for registration, transaction and record keeping — might be mined for better understanding of societal patterns, trends and policy impacts, or sanitised and released to fuel innovative products and services. A plethora of government reviews and initiatives have, especially over the last decade, led to the establishment of centres, networks and infrastructures to better understand societal phenomena using these data (Woollard 2014). Yet more recently, there has been a push to use administrative data to build models with the purpose of helping make day-to-day operational decisions in the management and delivery of public services, rather than providing general evidence to improve strategy or government-citizen interaction.

These new operational models are designed to serve as decision support or even to trigger automatic action. These are systems built primarily using *machine learning* techniques: algorithms which seek to identify patterns in datasets and render them into a usable form.

In this chapter, we focus on this trend, and attempt to answer elements of several important questions for the scholarship and practice of public administration:

1. What are the drivers and logics behind the use of machine learning in the public sector, and how should we understand it in the contexts of administrations and their tasks?

2. Is the use of machine learning in the public sector a smooth continuation of 'e-Government', or does it pose fundamentally different challenges to the practice of public administration?

3. How are public management decisions and practices at different levels enacted when machine learning solutions are implemented in the public sector?

We first explain the types of machine learning systems used in the public sector, detailing the processes and tasks that they aim to support. We then look at three levels of government — the macro, meso and the street-level — to map out, analyse and evaluate how machine learning in the public sector more broadly is framed and standardised across government considering unintended consequences (macro level), how it is designed and monitored in relation to proposed policy initiatives and existing public performance measurements, management and risk assessment mechanisms (meso level), and how it is implemented in the daily practices of frontline public service providers (micro level). We conclude that, while the use of machine learning in the public sector is mostly discussed with regard to its 'transformative effect' versus 'the dynamic conservatism' characteristic of public bureaucracies that embrace new technological developments (Hood 2008), it also raises several concerns about the skills, capacities, processes and practices that governments currently employ, the forms of which can have value-laden, political consequences.

Information technology is supposed to be a 'central force' to transformations in public management (Hood 2000 p. 17), although despite decades of promise of transformation, these tools usually fused onto existing practices rather than altering them at a deeper level (Margetts 1999). Scholars have argued that in recent years technologies have taken centre stage, and repositioned some of the trajectories of New Public Management into 'digital era governance'. They point to trends such as the digital re-integration of siloed services, data sharing practices aimed at creating a 'one-stop-shop' and 'end-to-end' service delivery with minimal repeated information gathering, and briefly discuss the rise of interest in 'zero touch technologies' (today often referred to as automated decision-making, following the EU Data Protection Directive 1995 and the later GDPR). This integration of digital technological developments in the public sector have led some to claim that 'technological change influences administrative capacity in public organisations', as the routines and capabilities of public organisations 'co-evolve with technology while being influenced by the wider institutional context (i.e. innovation systems)' (Lember et al. 2017). However, this raises the question whether information technologies and digital processes augment administrative capacity, leading to better management and delivery of public services?

Recently, machine learning powered, algorithmic tools[1] have garnered a great deal of attention due to public and private entities' drive to obtain value from large and increasingly well-structured datasets they hold, combined with a range of narrow discoveries in the machine learning field, primarily indicating that tasks we thought were computationally very difficult, such as certain types of image recognition or strategic board-game play, might be less so. In order to understand how governments increasingly administer by algorithm, we now distinguish between two main types of systems using machine learning for operational purposes: *automation systems* and *augmentation systems*.

1.1 Automation systems

Automation systems attempt to increase the quantity or efficiency of routine public sector operations through computation. Here, machine learning is used to enable the automation of tasks which have complicated elements but a straightforward and relatively objective outcome — such as triaging phone-calls or correspondence to the right points of contact. The incremental automation of rule-based processes is far from new, with public institutions such as tax agencies seeing it as an organisational ambition over many decades, with varying success (Margetts 1999). For processes that can be translated to rule-based systems with completeness and fidelity, progress continues at a slow-burn pace. Many barriers to rote automation surround classic challenges of legacy systems, as well as the slow and surprising creep of information technology in government over time, which has seen a greater fusion of data systems onto locked-in or slow-moving existing practices, rather than the transformative effect that had long been anticipated (cf. Downs 1967; Dunleavy et al. 2006). New technologies such as robotic process automation have already further aided integration by using computational techniques to automatically connect systems that do not naturally work together (Willcocks & Lacity 2016). Similarly, machine learning technologies provide improved tools, such as translation, image or handwriting recognition, which can be 'plugged in' to chains of automation for straightforward tasks. This follows the 'transformative vision' of information and communication technologies in the public sector, whereby technological innovations can lead to new 'government instrumentalities and operations', creating more effective ways of managing public portfolios, and more efficient and personalised public service delivery (Hood 2008).

Yet many administrative tasks are not straightforward and easily reduced or defined. Issues concerning operational decision-makers that might seem rote and 'objective' may be less so on closer inspection, and instead contain highly subjective and political aspects. Some researchers have historically pointed to a subset of tasks that therefore resist automation. An early empirical study of information systems in U.S. cities concluded that the

[1] In this work, 'algorithm' and its derivative terms are used as a shorthand for machine learning technologies and models, computational methods that, in practice, allow fine-grained patterns to be discerned in datasets and acted upon. This differs from a historical computer science definition of algorithms, which refers to repeatable computational processes, and which are not unfamiliar to government, business or practice.

political nature of some tasks, such as measuring internal departmental goals or deciding on external decisions (e.g. planning), may never allow them to be dramatically affected by computerisation—and that '[p]lanners and policy makers are especially cognizant of this reality' (Northrop et al. 1990 p. 512). 'Such models', they argued, 'would require criteria for defining problems and evaluating solutions, analysis of data in several files, and information that cannot be automated, such as interest group feelings about problems or support for various solutions.' Given the trend they saw to 'devalue community statistics and, instead, to emphasize the opinions of the affected citizens', they claimed 'it is likely that computerized information will have little impact on city planning decisions in the near future' (Northrop et al. 1990 p. 510). This sentiment has a longer history in public administration. Michael Lipsky, for example, claimed that 'the nature of service provision calls for human judgment that *cannot be programmed and for which machines cannot substitute*' (Lipsky 2010 p. 161) [emphasis added].

The implication is that equitable and effective public services require judgement that cannot be quantified, reduced or encoded in fully automated systems. These are issues familiar from the study of artificial intelligence and the law in the early nineties, when it became clear that the application of these systems led to grey zones of knowledge in problem-solving, and that, formally, codification was only effective in 'some highly specific, syntactically complex but semantically un-troubling domains' (Edwards & Veale 2017 p. 24). This in turn is connected to the indeterminacy of law: particularly the prevalence of terms with an 'open textured' nature, where the term's use or extension cannot be determined in advance of its application (Bench-Capon & Sergot 1988); where the connections between terms are vague in nature (Prakken 1997; Zeleznikow 2004); or where a series of factors are expected to be weighted and have relative importance assigned in a manner difficult to prescribe or render replicable (Christie 1986). At a larger, more strategic scale, the literature on the governance of sociotechnical problems has similarly emphasised the intractability of 'unstructured' or 'semi-structured' problems where there is a lack of consensus around appropriate means and/or ends, and how participatory processes that open up rather than close down are required to socially reach more navigable issues (Hoppe 2010).

Automation systems always bring politicised elements in the public sector, from encouraging the shifting and avoidance of blame, the increased rigidity of rules, and the types of 'edge cases' on which the systems will fail (Smith et al. 2010). They also serve to prioritise some public values, such as consistency and efficiency, above others (Hood 03/1991). However, where approaches with significant grey zones are automated, the value-laden nature of automation is accentuated, as the systems have to determine on which basis to make decisions within the grey zones of decision-making. This makes it necessary to ensure that automation systems, particularly ambitious ones, are well-encompassed by frameworks for suitable accountability.

Others have taken a different stance in relation to grey areas (Martinho-Truswell 2018), arguing that tasks that previously appeared to require human judgement, can now be *better*

decided upon with the help of statistical models such as machine learning systems. This leads to the second category: *augmentation systems*.

1.2 Augmentation systems

This second category of technological solutions, which we term *augmentation systems*, stems from a belief that machine learning does not just help cheapen or hasten decision-making, but can *improve it*.

What would it be to improve a decision? It is useful to point to a definition of machine learning from Mitchell (1997)[2]: we say that a machine learns when its *performance* at a certain *task* improves with *experience*. Here, performance, task and experience are captured through data, which are determined by designers. Improvement, or learning, can only be discussed once these three areas *at the very least* are accurately implemented. At a minimum, this requires that the aims of policy are quantifiable and quantified: a highly value-laden task in and of itself.

Traditionally, ensuring that policy is implemented with fidelity and legitimacy, and that public service delivery decisions are made in an equitable, effective and efficient manner, has fallen within the remit of 'bureaucratic professionalism', which itself carries tensions between responsiveness, as a means of enacting professional judgement, and standardised performance, as a means of ensuring best practice (Kearney & Sinha 1988; Stivers 1994). 'Bureaucratic professionalism' has itself changed from the Weberian model of administrative integrity and impartiality in the public interest, to (new) public management (Dahlström et al. 2011) and there has been growing recognition of its limitations (Bevan & Hood 2006; Dunleavy & Hood 1994; Hood & Peters 2004; Lapsley 2009). This shift has not only led to increased questioning of the effectiveness of measuring, standardising and auditing public sector performance for the public interest, but also brought about new conceptions of the role of the bureaucrat as negotiator and co-creator of public values with the citizens (J. V. Denhardt & Denhardt 2015; R. B. Denhardt & Denhardt 2000). In this respect, one could argue that this shift to new public service (NPS) is supporting the public servant's professional responsibility for more responsiveness in the management and delivery of public services, which augmentation systems could support.

Interestingly, in studies of digitisation of government (e.g. Dunleavy et al. 2006), there seems an almost complete omission of anticipation of the augmentative and predictive logics we have seen draw attention today. Programmes such as the *Integrated Data Infrastructure* in New Zealand (2016) have been designed not (just) for the purpose of creating 'one-stop shops' for accessing and delivering public services via interoperable, cross-departmental solutions (i.e. e-government 1.0 and 2.0), but for the purpose of 'informing decision-makers to help solve complex issues that affect us all, such as crime and vulnerable children'. Such programmes are thus established in order to augment the analytic and anti-

[2] This definition of machine learning has been popularised in the social, legal and philosophical studies of the technology by Mireille Hildebrandt, for example in Hildebrandt (2015).

cipatory capacity of contemporary governments 'to systematically use knowledge to inform a more forward-looking and society-changing style of policy-making' (Lodge & Wegrich 2014). These augmentations are hoped to help governments navigate coupled and complex problems that have ramifications outside the siloed organisational and decisional structures in which government departments still operate (i.e. wicked problems) (Andrews 2018).

The nature of the analytic capacity algorithmic augmentation systems are supposed to improve, particularly in the context of linked administrative data combined with additional data sources, is that it is possible to 'mine' data for insights public professionals alone would miss. In areas such as tax fraud detection, ambitions do not stay at replicating existing levels of success with reduced staff cost, but to do 'better than humans' (Milner & Berg 2017 p. 15). In highly value-charged areas where accuracy costs lives, such as child welfare and abuse, it is common to hear calls after a scandal that a tragedy 'could have been prevented', or that the 'information needed to stop this was there'.[3] Increased accuracy and the avoidance of human bias, rather than just the scalability and cost-efficiency of automation, is cited as a major driver for the development of machine learning models in high stakes spaces such as these (Cuccaro-Alamin et al. 2017).

In many ways, this logic continues the more quantified approach to risk and action found in the wide array of managerialist tools and practices associated with New Public Management. These have long had an algorithmic flavour, including performance measures and indicators, targets, and audits. Researchers have also emphasised that while these transformations are often justified as straightforward steps towards greater efficiency and effectiveness, in practice they represent core changes in expectations and in accountability (Burton & van den Broek 2009). Particularly in areas where professional judgement plays a key role in service delivery, such as social work, augmentation tools monitor and structure work to render individuals countable and accountable in new ways, taking organisations to new and more extreme bureaucratic heights of predictability, calculability and control.

Recently, these modern form of automation and augmentation tools have also received criticism from an emerging interdisciplinary field consisting of computer scientists, lawyers, sociologists of data, and more. It to this that we turn next.

1.3 Scholarly interest in socio-algorithmic issues

Scholars in the field of 'critical data studies' initially were fuelled by the belief that decisions made through the processing of large volumes of data have gained a veneer of neutrality (boyd 2016; boyd & Crawford 2012). Until recently, in high profile legal cases, the 'algorithmic neutrality' argument was commonly employed by large technology platforms to attempt to absolve themselves of responsibility (Kohl 2013). A plethora of studies have demonstrated how systems in areas such as natural language processing (Bolukbasi et al.

[3] Very rarely do those calling for this consider whether, were systems to be in place using such information, the number of false positives would be small enough to effectively enable identification of particular tragic cases.

2016; Caliskan et al. 2017), policing (Ensign et al. 2018), justice (Chouldechova 2017) and online content moderation (Binns et al. 2017), among others, are *far* from neutral. At a narrow level, they display detectable instances of unfairness between the way different groups are allocated results or represented in such systems (Crawford 2017). At a broader level, they raise deeper questions of whom the technology empowers or disempowers over time through its framing, context, purpose and use.

These concerns of algorithmic bias or discrimination (Barocas & Selbst 2016) have generated widespread interest in technical fields. In the computer science literature, it has spurred debate on how to mathematically specify fairness in such a way that it can be audited or placed as a statistical constraint during the training of a machine learning system (Kamiran et al. 2012). Such debate has spread from niche venues and conferences into the public eye (Courtland 2018), and is now a commonplace topic at the world's top research fora for machine learning, conferences such as the International Conference on Machine Learning (ICML) and Advances in Neural Information Processing Systems (NeurIPS).

Given that both automation and augmentation systems are seen as political, and at times, problematic, a range of speculative and analytical work has been undertaken around how systems can be held to account. This includes how assurance can be provided that the models actually used are the ones claimed to be used (Kilbertus et al. 2018; Kroll et al. 2016) or how the logic of systems can be inspected and explained (Edwards & Veale 2017) or how users feel about such explanations (Binns et al. 2018). Some argue that the potential for pervasive and adaptive deployment of these systems challenges the potential for them ever to be legitimate or meaningfully accountable (Yeung 2017a).

As questions of discrimination, power and accountability rise, there is a clear need to map both the problems and proposed means of managing them to institutions and practices in the public sector. Only a limited array of work has specifically focussed on administrative practices around these value-laden issues and technologies (e.g. Veale et al. 2018). In the following section, drawing on current and emerging policy initiatives, we will attempt to draw this link.

2 Emerging Management of Public Sector Machine Learning

We now outline and comment on the challenges of these systems that concern administrative structures at three different levels of public management and delivery. Firstly, the 'macro' scale, where governments play a strategic role in steering the implementation of this technology across public functions, and perform a duty to balance their benefits with their unintended consequences. Secondly, the 'meso' scale, where the delivery of individual public functions and policies is actualised in part through algorithmic systems and where the design, monitoring and evaluation of algorithmic systems is considered. Lastly, the 'micro' scale, looking at where specific frontline decisions affecting individuals, communities

and other entities are made and navigated by or with machine learning systems.

2.1 Macro: Government Strategy and Best Practices

At the highest level, governments have been discussing how to deal with machine learning and 'artificial intelligence' across many functions, including how to manage and promote innovation in the business environment, how to drive best practice and regulate misuse, and, most importantly, how to fit them into the day-to-day work of government. Here, we identify several domains where the integration of machine learning in the public sector appears to be occurring: the creation of new coordination mechanisms; new cross-cutting rights and obligations, and new capacities.

2.1.1 New coordination mechanisms

One way that governments have reacted to this change is by creating and strengthening coordinating bodies, codes and best practices that deal with algorithmic systems in the public sector.

New bodies and actors designed to take a cross-cutting role in data processing and analysis are emerging. Following a report from the Royal Society and British Academy suggesting a 'data stewardship body' (The Royal Society and the British Academy 2017), the UK Government is establishing a 'Centre for Data Ethics and Innovation' as an arms-length body (or 'quango') from the Department for Digital, Culture, Media and Sport. The exact terms of this body are, at the time of writing, out for consultation, but it is proposed that this centre 'support the government to enable safe and ethical innovation in the use of data and AI' through i) identifying steps to ensure that the law, regulation and guidance keep pace with developments in data-driven and AI-based technologies; ii) publishing recommendations to government on how it can support safe and ethical innovation in data and AI; and iii) providing expert advice and support to regulators (including for example the Information Commissioner's Office, Competition and Markets Authority and sector regulators) on the implications of data and AI uses and areas of potential harm. It is also proposed that such a body have a statutory footing (Department for Digital, Culture, Media and Sport 2018). Bodies with comparable cross-cutting competencies can be found emerging elsewhere, such as the French National Digital Council (*Conseil national du numérique*) and the recent German Data Ethics Commission (*Datenethikkommission*), designed to 'develop ethical guidelines for the protection of individuals, the preservation of the structure of social life and the safeguarding of prosperity in the information age' (Bundesministerium des Innern Für Bau Und Heimat 2018).

Alongside these bodies are emerging guidance documents and codes designed to inform the development of algorithmic systems across public functions. Perhaps the most complete at the time of writing is the UK Government's Data Ethics Framework (Department for Digital, Culture, Media & Sport 2018), currently in its second iteration. This document

is intended to guide 'the design of appropriate data use in government and the wider public sector', and is 'aimed at anyone working directly or indirectly with data in the public sector, including data practitioners (statisticians, analysts and data scientists), policymakers, operational staff and those helping produce data-informed insight'. Other countries are working on similar documents: for example, the Treasury Board of Canada is currently finalising a 'Directive on Automated Decision-Making' and further documentation on 'Responsible AI for the Government of Canada'. The city government of New York has legislated for a temporary 'task force' to produce a guidance report of a similar flavour.

While the documents above lack the binding force of law (although their use may be referenced as evidence of meeting broader obligations, such as the UK's public sector equality duty[4]), a few more binding documents have also been proposed. In the UK's Data Protection Bill 2018, a 'Framework for Data Processing in Government' was added by government amendment, which 'contains guidance about the processing of personal data' for the exercise of public functions. A person carrying out such a function must have regard to such a document when in force; as, controversially, should the independent Information Commissioner, if it appears 'relevant' to her. Such a document has not yet been published. It is also worth noting that the General Data Protection Regulation supports the creation and adoption of codes of conduct on a sectoral level, which may include codes of conduct relating to data management and public functions.

These documents and bodies may be useful to some of the low-stakes machine learning systems that are being built and implemented, to do tasks such as filter comments or triage parliamentary questions,[5] but may be less compelling for times where the stakes are higher or involve fundamental rights, liberties or 'taboo trade-offs' that are difficult to apply principles to (Fiske & Tetlock 1997). The persistently non-binding nature of these strategic documents and bodies, as well as the lack of resource of statutory obligation to consider their opinion or guidance, may seriously limit their impact. Experience from the wing-clipping of other digital cross-cutting bodies such as the UK's GDS (Hill 2018) would indicate that any binding effort in this area would be subject to bureaucratic competition and 'turf wars' between government departments (Bannister 2005), particularly exacerbated if data becomes the central force in government that many strategy documents hope.

2.1.2 New cross-cutting rights and obligations

Cross-cutting individual rights have been positioned as a proportionate counterbalance to public sector algorithmic systems. Fully automated decisions in many parts of government have been under the remit of data protection law across Europe since 1995, yet the requirement under these provision for decisions to be *solely* based on automated processing before requirements or safeguard triggers has limited their application to augmentation systems,

[4] Equality Act 2010, s 149.

[5] See e.g. the UK Ministry of Justice's parliamentary questions tool https://github.com/moj-analytical-services/pq-tool or the Government Digital Service's comment clustering tool https://dataingovernment.blog.gov.uk/2016/11/09/understanding-more-from-user-feedback/.

which typically have a human mediating between the computer system and the decision, even nominally (Edwards & Veale 2017).[6]

In France, where the automated decision-making rights in data protection law originate from (Bygrave 2001), 2016 legislation provided for further individual rights concerning cross-cutting government decisions based on an 'algorithmic treatment'. These provisions, placed into the French Administrative Code in early 2017, specify that upon request, individuals are provided with:

1. the degree and the mode of contribution of the algorithmic processing to the decision making;
2. the data processed and its source;
3. the treatment parameters and, where appropriate, their weighting, applied to the situation of the person concerned; and
4. the operations carried out by the treatment

(following the translation in Edwards & Veale 2018).

All EU countries must lay down safeguards relating to fully automated, significant decisions taken on the basis of national law (Article 22, GDPR), but these will not necessarily include decision support systems within their remit, as these are not 'solely' automated. Of interest in the French provision is that only a 'degree' of contribution of the algorithm is required, bringing a range of systems, and potentially even information sources we would traditionally consider 'evidence', within scope (Edwards & Veale 2018).

Explanation rights seem popular, but there is a tension between explanation (that computer scientists may favour) and justification, which is a more classic feature of administrative decisions. While the 'treatment parameters' in French law seem clearly connected to a model, provisions elsewhere, such as the common law *duty to give reasons* in England and Wales, may only require some justification (potentially post-hoc) for an algorithmic decision rather than an inspection mechanisms at play (Oswald 2018).[7] It could be argued this is more appropriate — there is a difference between *explanation* in law and *justification*, and tracing back why a decision occurred (e.g. treatment parameters) may not serve to justify it (Hildebrandt 2017). On the other hand, justification does not allow analysis of errors within a system, which may be systematic or distributed in discriminatory ways across a population, and these errors cause real harm and time wasted, even if a court or dispute resolution service may overturn them.

[6] Whether augmentation systems produce *solely* automated decisions, and what organisational means might result in meaningful human oversight in individual cases, are discussed below in the section on the 'micro' scale of administration.

[7] The scope of this duty is limited however: 'courts [in England and Wales] have consistently avoided imposing any general duty in administrative law to give reasons for decisions' (Oswald 2018 p. 5).

2.1.3 New capacities

Another cross-cutting response has been to build capacities, training and career pathways in order to support machine learning in the public sector. The UK Government has created the 'Government Data Science Partnership', which is a collaboration between the Government Digital Service (GDS), Office for National Statistics (ONS) and the Government Office for Science. Under this brand, initiatives such as new career paths for civil servant 'data scientists', new capacity building schemes such as the 'Data Science Accelerator', where civil servants work on public sector projects with mentors, and a new national framework of skills for the roles falling within the 'Digital, Data and Technology' profession, have been proposed. In different parts of government, new capacity centres have been set up, such as the Home Office's Data Analytics Competency Centre (DACC, formerly HODAC) and the ONS's Data Science Campus.

Technology strategy and translation capabilities are also being introduced inside other parts of the public sector. Organisations such as the Centre for Data Ethics and Innovation, described above, also intend to hold capacity and capacity building functions according to the released consultation documents. The Information Commissioner's Office has launched a technology strategy with a specific focus on AI, machine learning and algorithmic systems, indicating its intention to hire post-doctoral fellows, establish technology sandboxes, administer collaborative research grant programmes to build knowledge, and engage specialist expertise in new ways (Information Commissioner's Office 2018).

Whether these will be sufficient to be effective is yet to be seen. In particular, capacities that are not deeply woven within the day-to-day policy process, or accessible to different projects, directorates or functions, may fail to help shape and define algorithmic interventions, particularly when decisions or engagement around vendors, project scoping and public procurement are made in different places and with different timescales and involvement. Furthermore, it is an open question as to whether these individuals, and their skills, remain in the public sector, particularly at a time of great demand for quantitative talent in all areas of economy and society.

2.2 Meso: From High Level Policy to Practice

The macro level described above attempts to make a horizontal, facilitating environment for algorithmic design, deployment and maintenance in the public sector. Zooming in towards specific policy instances provides a different view on the challenges and practices as they are developing on the ground. The 'meso' level, as we describe it, concerns the practices of implementing high-level policy intentions with algorithmic components into a deployable and deployed system. At this meso level of public programme design and implementation, one of the main challenges is how to best balance bureaucratic control over the quality of the algorithmic system with measurement, monitoring and reporting practices that show how decisions made about and by algorithmic systems meet the high-level guidelines and

requirements specified in binding or soft laws. Thus, the implementation of models and machine learning tools in the design and delivery of public programmes raises important questions about how public sector performance and the 'explicit outcomes of government action' are measured and evaluated, in line with the continued relevance of New Public Management in the public sector (Heinrich 2003; Behn 2005).

Implementation is an old theme in public administration. A canonical model, and one that has drawn the focus of much implementation research, assumes that policy intentions (e.g. the goals of an algorithm-assisted policy) are formulated at a high level, for example, in a policy document, and that these are then 'implemented' on-the-ground (see discussion in Hupe et al. 2014). Such implementation may be more or less faithful — or 'congruent' — to the higher level intentions (Hupe 2011). Bureaucrats between high-level policy-makers and the citizens may influence this congruence in a variety of ways (Bowen 1982).

Algorithmic systems clearly shake this classic notion of implementation with its focus on success and failure, as this assumes there is a vision of what 'the policy' looks like (Hupe et al. 2014). In practice, ambiguity is often inherent in policy, leaving lower levels to define it or further specify it, in ways which can be considerably political and value-laden (Knoepfel & Weidner 1982; Matland 1995). Ministers and senior civil servants do not specify the training parameters or data sources of algorithmic systems, and so ambiguity is practically unavoidable. Where are decisions for acceptable performance of fraud detection systems made — around false positives, false negatives, performance on demographics, interpretability of decisions, or so on? Many of these cannot even be made until a pilot system has been developed, implying (in an ideal world, perhaps) a ping-ponging process between implementation and 'political' decision-making which may not resemble current programme performance measurement and management practices. This is made even more difficult by statistical properties of these systems: some notions of fairness, for example, may be statistically incompatible with each other, even though they seem simultaneously desirable (Chouldechova 2017).

This situation gives significant discretionary, or at the very least agenda-setting power, to what Bovens and Zouridis (2002) call 'system-level' bureaucrats. Here, we contrast their roles to both higher level decision-makers such as ministers, as well as their contrasting point, the 'street-level' bureaucrats of Lipsky (Lipsky 2010 p. 12) (who we come to below in the 'micro' section).

2.2.1 Balancing quality assurance and accountability

The challenges of algorithmic systems within government still retain a frame of quality, rather than considering the subjectivity of the deployed systems. In the UK, the failure of a computational model resulted in the mis-awarding of a rail franchise (Comptroller And Auditor General 2012), and led to a new guidance handbook, the *Aqua Book* (HM Treasury 2015) which sets out standards for analytical modelling in terms of assuring 'quality analysis'. Other erratic algorithmic systems which have triggered policy failure can be viewed

through this frame. The Australian *Centrelink* welfare system made headlines and was subject to several parliamentary investigations when it sent thousands of individuals incorrect debt notices (Karp & Knaus 2018), leaving its own automated system far from compliant with the law (Carney 2018).

The guidance prescribed by the *Aqua Book* describes institutional quality assurance processes which increase in rigour and burden as models increase both in complexity and in business risk (Figure 1). These range from, at a basic level, ensuring version control, testing, and the following of guidelines, to periodic internal and external peer review or audit. Peer review for models is also emphasised within other documents, such the UK Government Data Ethics Framework, which states that "[p]eer review is an essential part of quality assurance. Get feedback from your own team or organisation's data science function. If you're working alone, you may need to look elsewhere to receive appropriate scrutiny".

Yet a focus on peer review however risks omitting what the recent critical algorithmic literature has been so vocal about — that even when a model seems to 'work', it may not 'work' for everyone. The Dutch civil service *does* have a similar set of guidelines which do recognise this issue, albeit not as explicitly as might be possible. In the late nineties, a senior statistician in the Netherlands National Institute for Public Health and the Environment (Rijksinstituut voor Volksgezondheid en Milieu, RIVM) published a national op-ed criticising the environmental assessment branch of relying too heavily on models rather than measurements, cautioning that this was a dangerous 'imaginary world' (Petersen 2012 pp. 1–2). For what might consider quite an arcane issue of policy analysis, this led to a great media outcry in the Netherlands, with headlines including 'Environmental institute lies and deceits', 'Kafka and the environmental numbers', 'Credibility crisis surrounding environmental numbers' and 'Society has a right to fair information, RIVM does not provide it' (van der Sluijs 2002). The issue was debated on the floor of the Parliament within a matter of days. As a result of this, guidance on the assessment and communication of uncertainty during modelling was drawn up (Petersen et al. 2013), designed for use by civil servants and modellers. It breaks down uncertainty by quality and location, explicitly including uncertainty that is rooted in values or value-laden outcomes, with complementary tools made available within the civil service for the undertaking of this analysis.

The approach taken in the Netherlands (which in practice is still used primarily for models of natural phenomena rather than models based on personal data, such as for social policy) is connected to ideas in 'post-normal science' regarding 'extended peer review', where the 'peer community is [..] extended beyond the direct producers, sponsors and users of the research, to include all with a stake in the product, the process, and its implications both local and global. This extension of the peer community may include investigative journalists, lawyers and pressure groups' (Funtowicz & Ravetz 1993). Where the value-laden stakes are high, and affected communities are found widely, then narrow peer review may not be suited for ensuring the legitimacy of public sector models (van der Sluijs 2002).

This notion of extended peer review is yet to gain traction in algorithmic practices within government, but seems more promising for the types of issues identified in the critical algorithm studies literature than the highly technical assumptions made by the *Aqua Book*. There are, however, movements in policy initiatives and proposals towards infrastructure which would enable such extended peer review. The UK Government Data Ethics Framework (Department for Digital, Culture, Media & Sport 2018) states that

> 'Developed data science tools should be made available for scrutiny wherever possible [...] Even if the model cannot be released publicly, you may be able to release metadata about the model on a continual basis, like its performance on certain datasets. If your data science application is very sensitive, you could arrange for selected external bodies, approved by your organisation, to examine the model itself in a controlled context to provide feedback. This could be expertise from another government department, academia or public body.'

The UK House of Commons Science and Technology Committee, in its report on algorithms in decision-making, has stated similar intentions to open up algorithmic systems to third party actors, writing that 'the Government should produce, publish, and maintain a list of where algorithms with significant impacts are being used within Central Government' and make them available for analysis, or if there are intellectual property issues, in a 'suitably de-sensitised format' (House of Commons Science and Technology Committee 2018). While un-credited in the Lords' report, a similar recommendation was present in the UK Government's review of quality assurance of analytic models in 2013 (HM Treasury 2013). This led to a published list of the models used in government, although the list primarily focuses on models of phenomena and policies rather than system that, for example, assess risk at the level of individuals or entities.[8] Third sector proposals for public sector organisations have similar aims. These include the Code of Standards for Public Sector Algorithmic Decision-Making proposed by Nesta (2018) ('Public sector organisations should publish details describing the data on which an algorithm was (or is continuously) trained, and the assumptions used in its creation, together with a risk assessment for mitigating potential biases') and the proposals around public sector machine learning from the AI Now Institute at NYU (Reisman et al. 2018), which states that there should be a 'comprehensive plan for giving external researchers and auditors meaningful, ongoing access to examine specific systems, to gain a fuller account of their workings, and to engage the public and affected communities in the process'. There have also been attempts by universities to establish a list of of 'newsworthy' algorithmic systems in a U.S. context, which bears similarities to the UK government's internal inventory effort described above.[9]

[8] See Annex D of HM Treasury (2013) at `https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/206948/review_of_govt_models_annex_d_table_of_returns.pdf`.

[9] See eg `http://algorithmtips.org/` and `https://www.muckrock.com/project/uncovering-algorithms-84/`.

These proposals are promising, however they typically focus narrowly on the models rather than the broader process. Extended peer review mechanisms that are limited to the software alone may be insufficient. While public sector organisations might be able to provide general information about the data used or model built in a transparent manner, either to the public or to third parties (Edwards & Veale 2018), it is unlikely that they will be able to transparently evidence the broader process, of which machine learning is only a part, through which policy options or prediction mechanisms were supported. Proposals for 'algorithmic transparency' often go beyond explaining individual actions of a model to call for information about intentions of the individuals and teams involved, and the environment a system was trained and tested in (Edwards & Veale 2017; Selbst & Barocas 2018). This seems sensible in a technical sense, as it is hard to detect issues such as bias and discrimination in models for both statistical as well as practical reasons (Veale & Binns 2017).

2.2.2 Monitoring algorithmic systems performance

One mode of achieving transparency that has been proposed is through the publication of broad and relevant algorithmic metadata. The idea of metadata for data with a social focus, considering issues such as bias and sampling has already been proposed in the form of 'datasheets' (Gebru et al. 2018), which cover procedural elements as well as substantive features of the datasets themselves. As proposed, these would include answers to questions such as why the dataset was created; who funded it; what preprocessing or cleaning was undertaken (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances); aspects concerning its collection (e.g. presence of consent and ability to revoke); and information on updating and maintenance.

Translating these sheets to models is more of a challenge. How exactly models should be tested is still a matter of debate, and is likely to be highly context specific. In some cases, it may be worth considering how well they perform compared to human judgement (both in terms of the quantity and quality of the errors made). In other situations, it might be ore important to see how they perform on different subsets of the population, and the segments concerned will in turn vary strongly. Given that models can fail in unexpected ways (such as being 'fooled', as the domain of adversarial machine learning shows; see Nguyen et al. (2014)), it is unlikely that all potential aspects of salience and interest could be preempted in any set of metadata. Furthermore, as the growth of algorithmic systems in the public sector continues, it may be the interaction effects of these systems in tandem with each other, rather than metadata in isolation, which becomes important in practice. Incorporating all the relevant features of the model is enough of a challenges; considering the myriad ways it will be trained, used and deployed in enough detail to judge for the purposes of responsibility or accountability appears more daunting still.

Interestingly, if transparency *were* meaningfully introduced into the broader processes surrounding algorithmic conception and implementation, this would shine a great deal

more light on policy development than implementation processes generally get. As more policy development involves algorithmic systems, this could even change the way that decisions in government are scrutinised more broadly. Freedom of Information law, at least in the UK, contains a notable exemption for the 'formulation or development of government policy', which is often engaged to provide a 'safe space' for decisions and trade-offs, and transparency rights extending to process are likely to push against this safe space.[10] Time will tell whether transparency provisions related to algorithmic systems substantially change the way governments deal with internal decision-making; we use this space to float the considerable possibility that they might.

Finally, whether this additional transparency would result in more public trust around these systems is questionable. Some research has shown that the effect of transparency on public legitimacy is moderated by policy area. In particular, decisions involving 'taboo trade-offs' (Fiske & Tetlock 1997; Tetlock et al. 2000), where, for example, human well-being is traded off against economic considerations, may be made to appear less legitimate when transparency is applied, at least in part due to greater salience placed on these trade-offs by doing so (de Fine Licht 2014). Examples of these might be in areas such as algorithmic detection of child abuse (see eg Chouldechova et al. 2018), where the social perception of the cost of false negatives is so high that it is likely to make discussion of appropriate levels of invasiveness, accuracy, cost difficult. Such transparency might seem positive, but in effect, might end up bringing unintended consequences at this meso level of algorithmic system design and monitoring. This is compounded by the quantitative nature of the deployed systems. While it might be possible to stomach that an office of humans could fail to spot a specific case of child abuse without knowing the crystallised details of the office's success rate, it seems somehow less palatable that a bureaucracy has decided to deploy a system with a 15% rate of false negatives; that is to say, there is an *expectation* that 15% of cases which turn out to merit investigation or intervention will not be noticed.

2.3 Micro: Frontline Service Delivery

After these systems are designed, with the specific practices at the meso level and the support from the macro level, they are deployed day-to-day at the micro level. The purposes these systems can be put to are varied. Current public sector machine learning applications include tax fraud detection; a range of uses in policing from detecting geographic areas, potential victims, or required staffing levels; the risk scoring of prisoners to deliver appropriate rehabilitative courses; and the use of pre-emptive tools for child protection (Veale et al. 2018). The use of a similar flavour of technologies, sometimes using rudimentary forms of machine learning (such as regression methods) for decision-making in government is far from new (Griffiths 2017). Much service delivery now works around more advanced

[10] This should be caveated by noting that the Information Commissioner states that this exemption does not apply to implementation, and it will ' therefore be important to identify where policy formulation or development ends and implementation begins' (Information Commissioner's Office 2016 para. 34). However, as we have seen, algorithmic systems make this difficult in practice.

'risk-based frameworks' (Yeung 2017b), where 'the development of decision-making frameworks and procedures to prioritise regulatory activities and the deployment of resources, principally inspection and enforcement activities, [are] organised around an assessment of the risks that regulated firms pose to the regulator's objectives' (Black 2005 p. 514). Regulators have long been overburdened: those managing inspections, for example, have rarely had the resources to meaningfully audit more than a fraction of the organisations governed. A risk-based approach historically promnised more effective uses of limited resources, while also serving to manage the operational risk of the organisation: that it will not meet its own goals.

2.3.1 Automated processing and the rights of the data subject

At this level—the level where individual decisions are made concerning citizens and entities—legal regimes and scholars have been concerned about the threats of such automation to procedural justice in both its substance (Crawford & Schultz 2014; Keats Citron & Pasquale 2014) and the perception of it by citizens (Binns et al. 2018). There are emerging legal norms and regimes which attempt to manage this by envisaging significant gatekeeping and oversight roles for front-line decision-makers. As we will describe, these appear in some ways to be in tension with the logics of New Public Management. We argue that this tension highlights dimensions and lessons of importance that have been considered in emerging public sector practices that some scholars have called 'New Public Service' (R. B. Denhardt & Denhardt 2000).

The emerging legal regimes we are discussing largely only trigger when decisions impact directly on individuals.[11] Intrinsic concerns around automation of decisions have been of particular concern in European law for decades, as some have argued that in contrast to the US, Europe has treated automated decisions of many flavours as an inherent threat to human dignity (Jones 2017). Administrative law, particularly in France, included provisions as early as 1978 forbidding automated decisions by default in the public sector, and giving individuals information rights when automated treatments were lawfully applied (Bygrave 2001). These were 'uploaded' to European level through Article 15 of the Data Protection Directive 1995 and Article 22 of the General Data Protection Regulation 2016, both similar provisions which forbid significant, automated decisions without a specific lawful basis (Edwards & Veale 2017). The decisions explicitly envisaged by these provisions include 'automatic refusal of an online credit application or e-recruiting practices without any human intervention', however, outside the law enforcement domain, and without explicit member state derogations in national law, this provision applies across the public sector, as well as to contractors that in practice are often engaged to manage systems such as welfare assessment and provision.

Similar to the provisions that preceded it in the 1995 Data Protection Directive, all such

[11] An ongoing debate exists around the extent to which such provisions do or should take into account the impact upon *groups*, rather than just individuals (Edwards & Veale 2017).

decisions require a legal basis before they are taken, one of explicit consent of the decision subject; necessity for the performance of a contract with the subject; or authorisation by national law. For the public sector, the latter will apply. In the UK, section 14 of the Data Protection Act 2018 mandates that the public body must 'as soon as reasonably practicable, notify the data subject in writing that a decision has been taken based solely on automated processing', and that 'the data subject may, before the end of the period of 1 month beginning with receipt of the notification, request the controller to— (i) reconsider the decision, or (ii) take a new decision that is not based solely on automated processing'.

For our purposes, it is important to note that whether these provisions apply depends on whether a decision is considered to be 'solely' based on automated processing. The group of data protection regulators that interpret the law across Europe, the Article 29 Working Party (now superseded by a new European body, the European Data Protection Board, which has vouched for this former guidance), state that for a decision to be considered 'not based solely on automated processing' as section 14(4)(b)(ii) requires, the individual using a system as decision-support must (Veale & Edwards 2018):

i) have authority and competence to challenge the decision, and

ii) routinely express disagreement with any decision-support tool they might be using, rather than just apply a generated profile.

Without these characteristics, such a decision will collapse into being considered solely automated, and require notification to the individual, an ability to challenge a decision (though a simple written request, rather than through an expensive and largely inaccessible route of judicial review), and meaningful information to be provided surrounding the logic of processing (Edwards & Veale 2017). In some cases, these may be seen as proportionate safeguards that should always be provided. Yet to provide these safeguards in all cases where decision-support systems are used seems extensive, and public sectors may also consider it a burden they do not wish to engage with. If government *is* to become more 'data-driven', it will inevitably want to avoid a situation where the use of any routine decision-support system triggers Article 22 of the GDPR.

2.3.2 Algorithmic decisions, legitimacy and frontline discretion

To avoid the triggering of this provision however presents an organisational challenge which pushes against some of the logics behind New Public Management. New Public Management emphasises the need for standardised procedures that can be easily evaluated across inputs, processes, outputs, outcomes and impacts, informed predominantly by market values of efficiency and tailored allocation of resources (Heinrich 2003). The Article 29 Working Party guidance, and this direction of law more generally, emphasises what scholars have called *organisational humanism*, a characteristic that would align more closely with the assumption of 'citizen engagement in the conduct of public administration' that the 'New Public Service' literature proposes. Those focussed on this approach have sought to replace the traditional top-down approach to authority in administrative structures with structures

more attentive to the needs and concerns of a variety of internal and external actors, in part to 'supplement or even replace the authority of role or status with the authority of knowledge or competence' (J. V. Denhardt & Denhardt 2007 pp. 36–37). Without imbuing the frontline of an organisation with the authority to challenge or change the systems imposed by higher levels, a great number of systems will be considered, legally, solely automated.

Yet organisational humanism presents certain accountability tensions. In particular, there is a tricky balance between the professional responsibility that this organisational structure promotes and the legitimacy of a more standardised approach.

While at the meso-level modelling process, as discussed, developers' value-laden decisions can be overseen through mechanisms such as extended peer review, at the micro-level decisions are too numerous and granular to augment accountability through third party oversight. Where this high level of autonomy is present, internal accountability mechanisms have usually emphasised professionalism—the practices and norms accepted by a domain (Romzek & Ingraham 2000). Yet in these 'decision factories' of service delivery (Bovens & Zouridis 2002), New Public Management has taken the trend away from professionalism and towards measurable standards amenable to market-based instruments of monitoring and incentivisation. Even in domains where professionalism remains, such as social work, NPM has led to a far greater standardisation of decisions through guides, codes and the like (Ponnert & Svensson 2016). Such standardisation might help at building legitimacy and accountability in other ways (such as treating similar individuals similarly), but also risks turning decision-support systems back into what would legally be considered 'solely' automated ones.

While it is early days for advanced, granular decision-support systems within frontline public entities, we should expect them to catalyse new hybrid forms of accountability within these decision factories. For example, might individual 'authority and competence', which brings with it risks of problematic or arbitrary discretion if not undertaken with expensive care and professionalism, be supplemented by low-cost group oversight, such as second opinions given by peers? How might the practices of individuals disagreeing with decisions be documented, and how will these lower-level choices interact with hierarchies that are likely to want to monitor and assess them? Some scholars have argued that decision-support systems can effectively hide discretionary activities from those seeking to monitor them (Jorna & Wagenaar 2007); even if managers feel they understand how discretionary authority is being used, limitations of monitoring systems in practice can serve to make these grey areas invisible on the ground (Buffat 2011, 2015). These issues may well create new approaches to monitoring and oversight which differ from those we see today, with their own transformative effects on frontline service delivery.

3 Concluding Remarks

The use of machine learning in the public sector appears to be growing in prevalence. Entities at all levels of government are increasingly using automation and augmentation systems to either increase the efficiency of public sector operations or to support public decision-making for complicated or complex policy issues and programmes. While, at this stage, the discussion about the use of algorithmic systems in the public sector is framed around the dichotomy between 'transformation' and 'dynamic conservatism' that Hood (2008) outlined with regard to most information and communication technologies used in and by government, algorithmic systems raise new concerns that are not captured in the e-government literature or, indeed, in the practice of New Public Management that governments still adhere to. Many of these concerns are highly political and value-laden, and thus deserve particular attention as these technologies expand the range of roles and domains they touch upon.

In this chapter, we reviewed and analysed the latest developments and challenges that governments who promote and deploy machine learning in the public sector are facing. Looking at the macro level of government strategy and guidelines, at the meso level of public programme design, monitoring and evaluation, and at the micro level of implementation in frontline service delivery, we argue that the deployment of machine learning in the public sector challenges established institutions and administrative practices. Above, we outline three main challenges:

1. *Macro level*: The creation of new cross-cutting individual rights and obligations that require new skills and administrative capacities in order to fully assess the intended and unintended consequences of machine learning on established public values of accuracy, fairness, transparency and equity;

2. *Meso level*: The development of more dynamic ways to measure, monitor and evaluate the inputs, process information, outputs, outcomes and impacts of public programmes using machine learning, which challenge established measures of public sector performance, quality and risk assessment;

3. *Micro level*: The emergence of new tensions between the legitimacy of algorithmic decisions used in frontline service delivery, the discretion of street-level bureaucrats when employing, assessing or overriding automated decisions, and the rights of the data subjects when these processes are used to inform the allocation of public goods and services, and the discretion that street-level bureaucrats.

Scholars of public administration and management should place themselves at the heart of debates around the deployment and strategies of these technologies at all levels. This requires a heightened connection, as we have drawn, between the emerging scholarship on the sociotechnical challenges of algorithmic systems and the work around digital technologies in the public sector. In this period of rapid change — not so much in technology as the

coping strategies and structures of administrators at all levels — studying these emerging processes alongside the computer scientists, sociologists, technology lawyers and anthropologists doing the same is likely to bring new questions to the fore that will be difficult for any single field to answer alone. Public management and public administration scholars can impart important contributions to this debate, but only in close collaboration and when being open to issues and challenges, particularly around discrimination and fairness, that may be currently be more salient in other fields than they traditionally have been in the study of topics such as risk regulation or e-Government.

4 Acknowledgements

Both authors are supported by the Engineering and Physical Sciences Research Council (EPSRC) under grants EP/M507970/1 [MV] and EP/N02334X/1 [IB].

References

Andrews, L. (2018). 'Public administration, public leadership and the construction of public value in the age of the algorithm and "Big Data"', *Public Administration*, 36: 397. DOI: 10.1111/padm.12534

Atkinson, K., Bench-Capon, T., & Modgil, S. (2006). 'Argumentation for Decision Support'. *Database and Expert Systems Applications*, pp. 822–31. Springer Berlin Heidelberg. DOI: 10.1007/11827405_80

Bamberger, K. A., & Mulligan, D. K. (2015). Privacy on the Ground: Driving Corporate Behavior in the United States and Europe. MIT Press.

Bannister, F. (2005). 'E-government and administrative power: the one-stop-shop meets the turf war', *Electronic Government*, 2/2. DOI: 10.1504/EG.2005.007092.

Barocas, S., & Selbst, A. D. (2016). 'Big Data's Disparate Impact', *California Law Review*, 104: 671–732. DOI: 10.15779/Z38BG31

Behn, R.D. (2003). 'Why Measure Performance? Different Purposes Require Different Measures', *Public Administration Review* 63/5, pp. 586–606. DOI: 10.1111/1540-6210.00322.

Bench-Capon, T., & Sergot, M. J. (1988). 'Towards a rule-based representation of open texture in law'. Walter C. (ed.) *Computer power and legal language*, pp. 39–61. Quorum Books: New York.

Bevan, G., & Hood, C. (2006). 'What's measured is what matters: Targets and gaming in the English public health care system', *Public Administration*, 84/3: 517–38. DOI: 10.1111/j.1467-9299.2006.00600.x

Binns, R. (2018). 'Fairness in Machine Learning: Lessons from Political Philosophy', Pro-

ceedings of the First Conference on Fairness, Accountability and Transparency (FAT*).

Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., & Shadbolt, N. (2018). "It's Reducing a Human Being to a Percentage'; Perceptions of Justice in Algorithmic Decisions', *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI'18)*. DOI: 10.1145/3173574.3173951

Binns, R., Veale, M., Van Kleek, M., & Shadbolt, N. (2017). 'Like trainer, like bot? Inheritance of bias in algorithmic content moderation'. Ciampaglia G. L., Mashhadi A., & Yasseri T. (eds) *Social Informatics: 9th International Conference, SocInfo 2017, Oxford, UK, September 13-15, 2017, Proceedings, Part II*, pp. 405–15. Springer International Publishing: Cham. DOI: 10.1007/978-3-319-67256-4_32

Black, J. (2005). 'The emergence of risk-based regulation and the new public risk management in the United Kingdom', *Public Law*, 512–49.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., & Kalai, A. (2016). 'Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings'. Lee D. D., Sugiyama M., Luxburg U. V., Guyon I., & Garnett. R. (eds) *Advances in Neural Information Processing Systems 29 (NIPS 2016)*.

Bovens, M., & Zouridis, S. (2002). 'From Street-Level to System-Level Bureaucracies: How Information and Communication Technology is Transforming Administrative Discretion and Constitutional Control', *Public Administration Review*, 62/2: 174–84. DOI: 10.1111/0033-3352.00168

Bowen, E. R. (1982). 'The Pressman-Wildavsky Paradox: Four Addenda or Why Models Based on Probability Theory Can Predict Implementation Success and Suggest Useful Tactical Advice for Implementers', *Journal of Public Policy*, 2/1: 1–21. Cambridge University Press. DOI: 10.1017/S0143814X0000176

boyd, d. (2016). 'Undoing the neutrality of Big Data', *Florida Law Review Forum*, 16: 226–32.

boyd, d., & Crawford, K. (2012). 'Critical questions for Big Data: Provocations for a cultural, technological, and scholarly phenomenon', *Information, Communication and Society*, 15/5: 662–79. DOI: 10.1080/1369118X.2012.678878

Buffat, A. (2011). Pouvoir discrétionnaire et redevabilité de la bureaucratie de guichet: Les taxateurs d'une caisse de chômage comme acteurs de mise en oeuvre. Université de Lausanne, Lausanne, Switzerland.

——. (2015). 'Street-level bureaucracy and e-government', *Public Management Review*, 17/1: 149–61. DOI: 10.1080/14719037.2013.771699

Bundesministerium des Innern Für Bau Und Heimat. (2018). 'Datenethikkommission'. *Government of Germany*. Retrieved August 4, 2018, from <https://www.bmi.bund.de/DE/themen/it-und-digitalpolitik/datenethikkommission/datenethikkommission-node.html>

Burrell, J. (2016). 'How the machine 'thinks': Understanding opacity in machine learning

algorithms', *Big Data & Society*, 3/1. DOI: 10.1177/2053951715622512

Burton, J., & van den Broek, D. (2009). 'Accountable and Countable: Information Management Systems and the Bureaucratization of Social Work', *British Journal of Social Work*, 39/7: 1326–42. DOI: 10.1093/bjsw/bcn027

Bygrave, L. A. (2001). 'Automated Profiling: Minding the Machine: Article 15 of the EC Data Protection Directive and Automated Profiling', *Computer Law & Security Review*, 17/1: 17–24. DOI: 10.1016/S0267-3649(01)00104-2

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). 'Semantics derived automatically from language corpora contain human-like biases', *Science*, 356/6334: 183–6. DOI: 10.1126/science.aal4230

Carney, T. (2018). 'The New Digital Future for Welfare: Debts Without Legal Proofs or Moral Authority?', *UNSW Law Journal Forum, Sydney Law School Research Paper No. 18/15*.

Chouldechova, A. (2017). 'Fair prediction with disparate impact: A study of bias in recidivism prediction instruments', *Big Data*, 5/2: 153–63. DOI: 10.1089/big.2016.0047

Chouldechova, A., Benavides-Prado, D., Fialko, O., & Vaithianathan, R. (2018). 'A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions'. Friedler S. A. & Wilson C. (eds) *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, Proceedings of Machine Learning Research, Vol. 81, pp. 134–48. New York, NY, USA.

Christie, G. C. (1986). 'An essay on discretion', *Duke LJ*. HeinOnline.

Comptroller and Auditor General. (2012). *Lessons from cancelling the InterCity West Coast franchise competition*. London: National Audit Office.

Courtland, R. (2018). 'Bias detectives: the researchers striving to make algorithms fair', *Nature*, 558/7710: 357. Nature Publishing Group. DOI: 10.1038/d41586-018-05469-3

Crawford, K. (2017). 'The Trouble with Bias'. *NIPS 2017 (Keynote)*.

Crawford, K., & Schultz, J. (2014). 'Big data and due process: Toward a framework to redress predictive privacy harms', *Boston College Law Review,* 55: 93–128.

Cuccaro-Alamin, S., Foust, R., Vaithianathan, R., & Putnam-Hornstein, E. (2017). 'Risk assessment and decision making in child protective services: Predictive risk modeling in context', *Children and Youth Services Review*, 79: 291–8. DOI: 10.1016/j.childyouth.2017.06.027

Dahlström, C., Lapuente, V., & Teorell, J. (2011). 'The Merit of Meritocratization: Politics, Bureaucracy, and the Institutional Deterrents of Corruption', *Political Research Quarterly*, 65/3: 656–68. SAGE Publications Inc. DOI: 10.1177/1065912911408109

Denhardt, J. V., & Denhardt, R. B. (2007). *The New Public Service: Serving, Not Steering*. M.E. Sharpe.

——. (2015). 'The New Public Service Revisited', *Public Administration Review*, 75/5: 664–72. DOI: 10.1111/puar.12347

Denhardt, R. B., & Denhardt, J. V. (2000). 'The New Public Service: Serving Rather than Steering', *Public Administration Review*, 60/6: 549–59. Wiley Online Library. DOI: 10.1111/0033-3352.00117

Department for Digital, Culture, Media and Sport. (2018). 'Centre for Data Ethics and Innovation Consultation'. *HM Government*. Retrieved April 8, 2018, from <https://www.gov.uk/government/consultations/consultation-on-the-centre-for-data-ethics-and-innovation/centre-for-data-ethics-and-innovation-consultation>

Department for Digital, Culture, Media & Sport. (2018). 'Data Ethics Framework'. *HM Government*. Retrieved June 22, 2018, from <https://www.gov.uk/government/publications/data-ethics-framework>

Downs, A. (1967). 'A Realistic Look at the Final Payoffs from Urban Data Systems', *Public Administration Review*, 27/3: 204–10. DOI: 10.2307/973283

Dunleavy, P., & Hood, C. (1994). 'From old public administration to new public management', *Public Money & Management*, 14/3: 9–16. Routledge. DOI: 10.1080/09540969409387823

Dunleavy, P., Margetts, H., Bastow, S., & Tinkler, J. (2006). *Digital Era Governance: IT Corporations, the State and e-Government*. Oxford: Oxford University Press. DOI: 10.1093/acprof:oso/9780199296194.001.0001

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). 'The role of trust in automation reliance', *International Journal of Human-Computer Studies*, 58/6: 697–718. DOI: 10.1016/S1071-5819(03)00038-7

Edwards, L., & Veale, M. (2017). 'Slave to the Algorithm? Why a 'Right to an Explanation' is Probably Not the Remedy You Are Looking For', *Duke Law and Technology Review*, 16/1: 18–84. DOI: 10.2139/ssrn.2972855

——. (2018). 'Enslaving the algorithm: From a 'right to an explanation' to a 'right to better decisions'?', *IEEE Security & Privacy*, 16/3: 46–54. DOI: 10.2139/ssrn.3052831

Ensign, D., Friedler, S. A., Neville, S., Scheidegger, C., & Venkatasubramanian, S. (2018). 'Runaway Feedback Loops in Predictive Policing'. *Proceedings of the 1st Conference on Fairness, Accountability, and Transparency (FAT*)*.

Erickson, P., Klein, J. L., Daston, L., Lemov, R., Sturm, T., & Gordin, M. D. (2013). *How Reason Almost Lost Its Mind: The Strange Career of Cold War Rationality*. University of Chicago Press.

de Fine Licht, J. (2014). 'Policy Area as a Potential Moderator of Transparency Effects: An Experiment', *Public Administration Review*, 74/3: 361–71. DOI: 10.1111/puar.12194

Fiske, A. P., & Tetlock, P. E. (1997). 'Taboo Trade-offs: Reactions to Transactions That Transgress the Spheres of Justice', *Political Psychology*, 18/2: 255–97. DOI: 10.1111/0162-

895X.00058

Funtowicz, S. O., & Ravetz, J. R. (1993). 'Science for the post-normal age', *Futures*, 25/7: 739–55. Elsevier. DOI: 10.1016/0016-3287(93)90022-L

Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2013). 'A survey on concept drift adaptation', *ACM Computing Surveys*, 1/1. DOI: 10.1145/2523813

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2018). 'Datasheets for Datasets'. *Presented at the 5th Workshop on Fairness, Accountability, and Transparency in Machine Learning (FAT/ML), Stockholm, Sweden*.

Grgić-Hlača, N., Zafar, M. B., Gummadi, K. P., & Weller, A. (2017). *On Fairness, Diversity and Randomness in Algorithmic Decision Making. arXiv [stat.ML]*. Retrieved from arXiv.

Griffiths, A. (2017). Forecasting Failure: Assessing Risks to Quality Assurance in Higher Education Using Machine Learning (PhD). King's College London.

Heinrich, C.J. (2003). 'Measuring public sector performance and effectiveness', Peters, B.G. & Pierre, J. (eds.), *The SAGE Handbook of Public Administration,* SAGE, pp.24-38.

Hildebrandt, M. (2015). *Smart technologies and the end(s) of law*. Cheltenham, UK: Edward Elgar.

——. (2017). 'Privacy As Protection of the Incomputable Self: From Agnostic to Agonistic Machine Learning', DOI: 10.2139/ssrn.3081776

Hill, R. (6 Jul 2018). '"Toxic" Whitehall power culture fingered for GDS's fall from grace'. *The Register*.

HM Treasury. (2013). Review of quality assurance of government models. London: HM Government.

——. (2015). The Aqua Book: guidance on producing quality analysis for government. London: HM Government.

Hood, C. (03/1991). 'A Public Management for All Seasons?', *Public Administration*, 69/1: 3–19. DOI: 10/bdwbfj

——. (2000). The art of the state: Culture, rhetoric, and public management. Oxford University Press._—

——. (2008) 'The Tools of Government in the Information Age', Goodin, R.E., Moran, M., & Rein, M. (eds.) *The Oxford Handbook of Public Policy*. Oxford University Press.

Hood, C., & Peters, G. (2004). 'The Middle Aging of New Public Management: Into the Age of Paradox?', *Journal of Public Administration Research and Theory*, 14/3: 267–82. DOI: 10.1093/jopart/muh019

Hoppe, R. (2010). The Governance of Problems: Puzzling, Powering and Participation. Bristol: Policy Press.

House of Commons Science and Technology Committee. (2018). *Algorithms in decision-*

making (HC 351). London: UK Parliament.

Hupe, P. L. (2011). 'The Thesis of Incongruent Implementation: Revisiting Pressman and Wildavsky', *Public Policy and Administration*, 26/1: 63–80. DOI: 10/dp9xr9

Hupe, P. L., Hill, M., & Nangia, M. (2014). 'Studying implementation beyond deficit analysis: The top-down view reconsidered', *Public Policy and Administration*, 29/2: 145–63. DOI: 10.1177/0952076713517520

Information Commissioner's Office. (2016). *Government policy (section 35): Freedom of Information Act*. Wilmslow: ICO.

——. (2018). *Technology Strategy, 2018-2021*. Wilmslow, UK: ICO.

Jones, M. L. (2017). 'The right to a human in the loop: Political constructions of computer automation and personhood', *Social Studies of Science*, 47/2: 216–39. DOI: 10.1177/0306312717699716

Jorna, F., & Wagenaar, P. (2007). 'The 'iron cage' strengthened? Discretion and digital discipline', *Public Administration*, 85/1: 189–214. DOI: 10.1111/j.1467-9299.2007.00640.x

Kamarinou, D., Millard, C., & Singh, J. (2016). 'Machine Learning with Personal Data',. papers.ssrn.com.

Kamiran, F., Calders, T., & Pechenizkiy, M. (2012). 'Techniques for discrimination-free predictive models'. Custers B., Calders T., Schermer B., & Zarsky T. (eds) *Discrimination and Privacy in the Information Society*, pp. 223–40. Springer: Berlin, Heidelberg.

Karp, P., & Knaus, C. (2018). 'Centrelink robo-debt program accused of enforcing "illegal" debts'. *The Guardian*.

Kearney, R. C., & Sinha, C. (1988). 'Professionalism and bureaucratic responsiveness: Conflict or compatibility?', *Public Administration Review*, 571–9. JSTOR.

Keats Citron, D., & Pasquale, F. (2014). 'The scored society: Due process for automated predictions', *Washington Law Review* , 89/1: 1–33.

Kilbertus, N., Gascon, A., Kusner, M., Veale, M., Gummadi, K., & Weller, A. (2018). 'Blind Justice: Fairness with Encrypted Sensitive Attributes', *Proceedings of the 35th International Conference on Machine Learning (ICML 2018)*. International Machine Learning Society (IMLS).

Knoepfel, P., & Weidner, H. (1982). 'Formulation and Implementation of Air Quality Control Programmes: Patterns of Interest Consideration', *Policy & Politics*, 10/1: 85–109. DOI: 10.1332/030557382782628914

Kohl, U. (2013). 'Google: the rise and rise of online intermediaries in the governance of the Internet and beyond (Part 2)', *International Journal of Law and Information Technology*, 21/2: 187–234. DOI: 10.1093/ijlit/eat004

Kroll, J. A., Huey, J., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., & Yu, H. (2016). 'Accountable Algorithms', *University of Pennsylvania Law Review*, 165.

Lapsley, I. (2009). 'New Public Management: The Cruellest Invention of the Human Spirit? 1', *Abacus*, 45/1: 1–21. Wiley Online Library. DOI: 10.1111/j.1467-6281.2009.00275.x

Lember, V., Kattel, R., & Tõnurist, P. (2017) Technological Capacity in the Public Sector: The Case of Estonia. *IIPP Working Paper Series, 2017-03*.

Lipsky, M. (2010). Street-level bureaucracy: Dilemmas of the individual in public services. New York: Russell Sage Foundation.

Lodge, M., & Wegrich, K. (2014). The Problem-solving Capacity of the Modern State: Governance Challenges and Administrative Capacities. Oxford University Press.

Margetts, H. (1999). Information Technology in Government: Britain and America. London: Routledge.

Martinho-Truswell, E. (2018). 'How AI Could Help the Public Sector'. *Harvard Business Review*.

Matland, R. E. (1995). 'Synthesizing the Implementation Literature: The Ambiguity-Conflict Model of Policy Implementation', *Journal of Public Administration Research and Theory*, 5/2: 145–74. Oxford University Press. DOI: 10.1093/oxfordjournals.jpart.a037242

Milner, C., & Berg, B. (2017). *Tax Analytics: Artificial Intelligence and Machine Learning*. PwC Advanced Tax Analytics & Innovation.

Mitchell, T. M. (1997). *Machine learning*. Burr Hill, IL: McGraw Hill.

Nguyen, A., Yosinski, J., & Clune, J. (2014). 'Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images', *arXiv:1412.1897 [cs]*.

Northrop, A., Kraemer, K. L., Dunkle, D., & King, J. L. (1990). 'Payoffs from Computerization: Lessons over Time', *Public Administration Review*, 50/5: 505–14. [American Society for Public Administration, Wiley]. DOI: 10.2307/976781

Oswald, M. (2018). 'Algorithm-assisted decision-making in the public sector: framing the issues using administrative law rules governing discretionary power', *Philosophical Transactions of the Royal Society of London. Series A: Mathematical and Physical Sciences*, 376/2128: 20170359. The Royal Society. DOI: 10.1098/rsta.2017.0359

Petersen, A. C. (2012). Simulating nature: A philosophical study of computer-simulation uncertainties and their role in climate science and policy advice. London: CRC Press.

Petersen, A. C., Janssen, P. H. M., van der Sluijs, J. P., Risbet, J. S., Ravetz, J. R., Arjan Wardekker, J., & Martison Hughes, H. (2013). *Guidance for uncertainty assessment and communication*. The Hague, NL: PBL Netherlands Environmental Assessment Bureau.

Ponnert, L., & Svensson, K. (2016). 'Standardisation—the end of professional discretion?', *European Journal of Social Work*, 19/3-4: 586–99. Routledge. DOI: 10.1080/13691457.2015.1074551

Prakken, H. (1997). Logical tools for modelling legal argument. Dordrecht: Kluwer.

Reisman, D., Schultz, J., Crawford, K., & Whittaker, M. (2018). *Algorithmic Impact As-*

sessments: A Practical Framework for Public Agency Accountability. New York: AI NOW Institute.

Romzek, B. S., & Ingraham, P. W. (2000). 'Cross Pressures of Accountability: Initiative, Command, and Failure in the Ron Brown Plane Crash', *Public Administration Review*, 60/3: 240–53. DOI: 10.1111/0033-3352.00084

Seaver, N. (2013). 'Knowing algorithms'. Paper presented at Media in Transition 8, Cambridge, MA.

——. (2017). 'Algorithms as culture: Some tactics for the ethnography of algorithmic systems', *Big Data & Society*, 4/2. DOI: 10.1177/2053951717738104

Selbst, A., & Barocas, S. (2018). 'The Intuitive Appeal of Explainable Machines', draft available on SSRN.

Skitka, L. J., Mosier, K. L., & Burdick, M. (1999). 'Does automation bias decision-making?', *International Journal of Human-Computer Studies*, 51: 991–1006. DOI: 10.1006/ijhc.1999.0252

van der Sluijs, J. P. (2002). 'A way out of the credibility crisis of models used in integrated environmental assessment', *Futures*, 34/2: 133–46. DOI: 10.1016/S0016-3287(01)00051-9

Smith, M. L., Noorman, M. E., & Martin, A. K. (2010). 'Automating the public sector and organizing accountabilities', *Communications of the Association for Information Systems*, 26/1: 1.

Snellen, I. (2002). 'Electronic Governance: Implications for Citizens, Politicians and Public Servants', *International Review of Administrative Sciences*, 68/2: 183–98. SAGE Publications Ltd. DOI: 10.1177/0020852302682002

Stivers, C. (1994). 'The listening bureaucrat: Responsiveness in public administration', *Public Administration Review*, 364–9.

Sunstein, C. R. (2002). *Risk and Reason: Safety, Law, and the Environment*. Cambridge University Press.

Tetlock, P. E., Kristel, O. V., Elson, S. B., Green, M. C., & Lerner, J. S. (2000). 'The psychology of the unthinkable: taboo trade-offs, forbidden base rates, and heretical counterfactuals', *Journal of Personality and Social Psychology*, 78/5: 853–70.

The Royal Society and the British Academy. (2017). *Data management and use: Governance in the 21st Century*. London: The Royal Society and the British Academy.

Veale, M., & Binns, R. (2017). 'Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data', *Big Data & Society*, 4/2. DOI: 10.1177/2053951717743530

Veale, M., & Edwards, L. (2018). 'Clarity, Surprises, and Further Questions in the Article 29 Working Party Draft Guidance on Automated Decision-Making and Profiling', *Computer Law & Security Review*, 34/2: 398–404. DOI: 10.1016/j.clsr.2017.12.002

Veale, M., Van Kleek, M., & Binns, R. (2018). 'Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making', *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI'18)*. DOI: 10.1145/3173574.3174014

Willcocks, L. P., & Lacity, M. (2016). *Service automation robots and the future of work.*, p. 304. Ashford, UK: SB Publishing.

Woollard, M. (2014). 'Administrative Data: Problems and Benefits. A perspective from the United Kingdom'. Duşa A., Nelle D., Stock G., & Wagner G. G. (eds) *Facing the Future: European Research Infrastructures for the Humanities and Social Sciences*, pp. 49–60. SCIVERO Verlag: Berlin.

Yeung, K. (2017a). '"Hypernudge": Big Data as a mode of regulation by design', *Information, Communication and Society*, 20/1: 118–36. Routledge. DOI: 10.1080/1369118X.2016.1186713

——. (2017b). 'Algorithmic regulation: A critical interrogation', *Regulation & Governance*, 347: 509. DOI: 10.1111/rego.12158

Zeleznikow, J. (2004). 'The Split-up project: induction, context and knowledge discovery in law', *Law, Probability and Risk*, 3/2: 147–68. DOI: 10.1093/lpr/3.2.147

Žliobaitė, I., Pechenizkiy, M., & Gama, J. (2016). 'An Overview of Concept Drift Applications'. Japkowicz N. & Stefanowski J. (eds) *Big Data Analysis: New Algorithms for a New Society*, Studies in Big Data, pp. 91–114. Springer International Publishing. DOI: 10.1007/978-3-319-26989-4_4