Information Polity 25 (2020) 471–490
DOI 10.3233/IP-200249
IOS Press

Administration by algorithm: A risk management framework

Frank Bannister[a,*] and Regina Connolly[b]
[a]*Trinity College Dublin, Dublin, Ireland*
[b]*Dublin City University, Dublin, Ireland*

Abstract. Algorithmic decision-making is neither a recent phenomenon nor one necessarily associated with artificial intelligence (AI), though advances in AI are increasingly resulting in what were heretofore human decisions being taken over by, or becoming dependent on, algorithms and technologies like machine learning. Such developments promise many potential benefits, but are not without certain risks. These risks are not always well understood. It is not just a question of machines making mistakes; it is the embedding of values, biases and prejudices in software which can discriminate against both individuals and groups in society. Such biases are often hard either to detect or prove, particularly where there are problems with transparency and accountability and where such systems are outsourced to the private sector. Consequently, being able to detect and categorise these risks is essential in order to develop a systematic and calibrated response. This paper proposes a simple taxonomy of decision-making algorithms in the public sector and uses this to build a risk management framework with a number of components including an accountability structure and regulatory governance. This framework is designed to assist scholars and practitioners interested in ensuring structured accountability and legal regulation of AI in the public sphere.

Keywords: Algorithm, artificial intelligence, e-government, governance, risk management

1. Introduction

This paper is concerned with a specific challenge facing contemporary society namely managing the risks that arise from the use of computer systems for supporting or making decisions in the public sector. This is not a new risk, but it is one that is being amplified by the increasing use of artificial intelligence (AI) in such decision-making. However, while much of the focus in this paper is on AI, we use the term "algorithm" rather than AI. This requires a brief explanation. To do this it is first useful to set out what is understood by AI and then briefly discuss the nature of algorithms.

Artificial intelligence is a broad church encompassing, in the view of some, everything from spell checkers to the self-aware computers of science fiction such as HAL in *2001 – A Space Odyssey* or 3CPO in *Star Wars* to name but two. The EU High Level Expert Group on Artificial Intelligence (2019) has sought to capture that breadth, proposing the following definition:

> "Artificial intelligence (AI) systems are software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the

*Corresponding author: Frank Bannister, Trinity College Dublin, Dublin, Ireland. E-mail: Frank.Bannister@tcd.ie.

best action(s) to take to achieve the given goal. AI systems can either use symbolic rules or learn a numeric model, and they can also adapt their behaviour by analysing how the environment is affected by their previous actions.
As a scientific discipline, AI includes several approaches and techniques, such as machine learning (of which deep learning and reinforcement learning are specific examples), machine reasoning (which includes planning, scheduling, knowledge representation and reasoning, search, and optimization), and robotics (which includes control, perception, sensors and actuators, as well as the integration of all other techniques into cyber-physical systems)." (EUHLEG 2019, p. 5)

A notable feature of AI is that its impact has been steadily spreading and deepening over a long period of time. Some technologies such as, say, social networking or the World Wide Web, emerged and became an important part of our social fabric in less than a decade. Others, such as the Internet (invented in 1969) and AI have been burrowing their way into people's lives for 50 years or more. Many developments in so-called weak AI such as expert systems or voice recognition, both of which were developed and improved over a long period, were once crude and unreliable, but are now sophisticated and widely adopted. More recent developments include facial recognition and machine learning.

Not all algorithms qualify as forms of AI. An algorithm is nothing more than series of steps designed to achieve a particular outcome. It is neither necessarily mathematical nor computer based. Amongst their many applications, algorithms are used to solve problems, to optimise, to carry out daily tasks and to make predictions. However, almost all non-trivial algorithms involve choices/decision-making at some level. From deciding when to lower the flaps on an aircraft to determining whose tax returns should be audited, computer systems across the world make vast numbers of decisions every minute. Almost all of these decisions are uncontroversial and/or do not have life changing implications for individuals. But some do. Of particular interest are complex and difficult to understand algorithms that make or recommend decisions that have material implications or even life-changing consequences for individuals, groups and/or societies. The danger of faulty algorithms was vividly and tragically illustrated by the crashes of two Boeing 737 800 Max aircraft in 2018 and 2019. Much less tragic errors, but errors with other forms of undesirable consequences for individual citizens, happen every day as a result of problems embedded in software. What is possibly even more worrying is that some of these problems may not be bugs, but features, i.e. policies and/or prejudices consciously or unconsciously built into the code and/or problems caused by biases in the underlying data that they use (Ntoutsi et al., 2020).

Consequently, while, as AI based algorithmic decision-making becomes more and more widespread, the focus of concern going forward has to be on this technology, it is necessary to consider all decision-making algorithms. Algorithms exist on a risk spectrum; they range from the insignificant to the momentous. We need to learn how to recognise, for any given algorithmic decision-making process, not just its benefits, but its potential risks and how to calibrate and manage such risks. We need to learn how to control algorithms before they come to control us any more than they already do. This paper will propose structures and mechanisms which can help ensure that we remain in control. While the word "algorithm" is the word most often used in this paper, the term "artificial intelligence" or AI will be used throughout this paper to refer specifically to its application to decision-making of the type being examined.

2. Literature review

2.1. Problems with AI and decision making

Concern about the potentially negative impacts of artificial intelligence have long been a staple of science fiction, both in literature and in the cinema. Many of these works have been prescient. To take

just one example, Primo Levi (1966) in one of his amusing short science fiction stories, *Pieno Impeigo* (Full Employment), describes the development of a machine that today we would call a search engine and briefly speculates on its implications. Academic research on this topic has, however, been more scarce largely because, while writers of fiction can indulge freely in speculation about the future, scholars have to be more measured – particularly when it comes to predicting the impacts of a new technology and the risks that such technologies may bring with them.

One of the first researchers to conceptualise risk in terms of its dimensions was Cunningham (1967). The distinct dimensions that he proposed were uncertainty and consequences – concepts today widely used in the form of expected outcomes. Both of these dimensions are salient when considering AI, as the many uncertainties regarding the potential impact AI on society are increasingly juxtaposed with an awareness that the consequences of unbridled AI could wreak havoc on individuals and on society more generally (see, for example, Bostrom, 2014). Away from such dystopian discourses, there are many practical risks arising from the use of AI in day to day operations that need to be managed (as would be the case with any new technology). As will be shown below, there is growing evidence of AI-related damage occurring, albeit for now only in a localised way. Rather than assuming a perspective that enumerates only potential positive or only negative AI scenarios and outcomes, a more realistic position is to acknowledge the uncertainty that exists and conceptualise the risks of AI as a set of potential trade-offs that need to be carefully evaluated and managed.

The work of Floridi et al. (2018) is particularly valuable in this regard. These authors identify four types of risk associated with AI that are relevant to society and consider how we can balance the opportunities that AI offers against the potential risks that it presents.

The first relates to the *enabling of human self-realisation without devaluing human abilities*. Floridi et al. point out that although smart automation may free us from the tedium of monotonous tasks (just as did invention of the washing machine) and allow us to spend our time more intelligently, the risks that arise are not just those of making older skills obsolete with the emergence of new ones, but rather the pace at which this is happening and the resulting unequal distribution of costs and benefits. For example, employment serves economic, social and even cultural purposes in our society and the devaluation of old skills has the potential to disrupt the nature of employment at the level of both the individual and society. Beyond its obvious economic implications, jobs are central to an individual's sense of personal identity and meaning, their self-esteem and their sense of self, whilst at a societal level de-skilling has potential to result in loss of human abilities and skills, thereby creating vulnerabilities in the event of AI malfunction. In this context, an extensive body of research has examined the effects of automation on the labour market (e.g. Manyika & Bughin, 2018; Acemoglu & Restrepo, 2017; Arntz et al., 2016; Karoly & Panis 2004), but much of that work has focused on robotic systems with limited intelligence which follow pre-programmed instructions to complete specific manual tasks. Emerging AI systems in contrast do not require such specific, bounded guidance and are capable of acting independently on insights obtained from data.

The second risk that Floridi et al. identify, one that is of particular relevance here, relates to the need to *ensure enhancement of human agency without removing human responsibility*. AI augments our intelligence in the same way as machines have previously augmented our abilities and outputs. However, they point to the need for responsibility in deciding what sort of AI we develop, how it is used and whether it should be shared with everyone. The question becomes to what degree do we abdicate responsibility for decision making to machines? The use of AI in parole decisions or creditworthiness decisions are just two of hundreds of possible examples. This question has a number of consequences that are discussed in subsequent sections.

In theory, an AI decision-making system can be objective or neutral (assuming that we can reach a consensus on what those terms mean – not a given by any means), but in most social decision-making contexts AI is almost never neutral and decisions made by complex algorithms are, almost by definition, never objective. To enable an algorithm to make a decision, values must be embedded in its code or indirectly in the data that it uses. Consequently, the use of AI in, for example, surveillance and military applications as well as use in social media and election result manipulation raises understandable concerns. As Manyika and Bughin (2018) note, AI has the potential to introduce new biases or perpetuate and institutionalise existing social and procedural biases. A recent report from the *AI Now* Institute (Myers West et al., 2019) found that the AI field, which is overwhelmingly white and male, is at risk of replicating historical biases and power imbalances. Because of its capabilities, AI is also at risk of amplifying biases and discriminatory practices. Examples of this abound, including use of image recognition services that make offensive classifications of minorities (as happened with *Google Photos*), chatbots that adopt hate speech (such as happened with Microsoft's Twitter bot *Tay* and systems like Amazon's facial recognition technology *Rekognition* which was marketed to law enforcement agencies, but has been shown to fail to recognise the gender of individuals with darker skin colours. According to Singer (2019) failures of this nature raise fundamental questions regarding whether facial recognition technology is suitable for use in important decision-making situations including job interviews, in drones and in police body cameras. Designed correctly, AI can amplify individual agency and strengthen moral systems, but as the Floridi et al. note, the degree of agency that we delegate to autonomous system is not a zero-sum game either pragmatically or ethically and it would be particularly unwise to ignore this fact.

The third risk they cite relates to the need to *increase societal capabilities without reducing human control*. AI presents a multitude of possibilities for augmenting and enhancing the collective capabilities of humans, assisting us to find solutions to enduring problems with significant individual and social impact, including early identification or even prevention of disease. But these authors note that it is precisely because of this, the fact that such technologies have the potential to be so powerful and disruptive, that they introduce proportionate risks. They argue that if we rely on AI to augment our own abilities in the wrong way, we may delegate important decisions to autonomous systems, systems that should remain at least partly subject to human supervision and choice. Worse, these potential harms could accumulate and become entrenched as more functions are delegated to machines. There is therefore a need to strike a balance between pursuing the opportunities offered by AI to improve human life on one hand and on the other hand ensuring that we remain in control of these developments and their effects

The final risk identified by Floridi et al. (2018) is *how we can cultivate social cohesion without eroding human self-determination*? In the face of current global societal problems such as climate change, terrorism, antimicrobial resistance and other similar problems, there is a need for high degrees of co-ordination between the stakeholders if these problems are to be successfully addressed.[1] AI has obvious unique abilities to support such coordination complexity and to support greater collaboration and a more cohesive response, through reminder notifications, identifying those who are transgressing socially agreed norms or standards etc. The emergence of social scoring in China is a particularly worrying development (Liang et al., 2018; Creemers, 2018). This can already result in people with low scores being denied certain services such as access to housing (Mitchell & Diamond, 2018).[2] As we seek to design societal frameworks that work for all, there are likely to be algorithmic systems that we adopt to self-nudge us to

[1] This article is being written during the middle of COVID-19 pandemic, but we have resisted the temptation to bring pandemics into the discussion.

[2] An episode of the TV series Black Mirror (Nosedive) explored this theme.

behave in socially preferable ways. However in adopting such systems, we run the risk of colluding with AI in eroding human self-determination as systems change human behaviour to accommodate outcomes. This is not, of course, a public sector monopoly. As Zubhoff (2019) has shown, private companies are seeking the same goal, i.e. the manipulation of humans and effectively a reduction in their agency.

2.2. AI-augmented government

Eggers and Beyer (2019) suggest that AI has potential to bring significant improvements in efficiency or cost savings to government by replacing existing decision-making processes and as the basis for policy determinations. The rationale behind this view is that government agencies are characterised by voluminous data and AI can both process and learn from these data much more efficiently than humans, thereby making it possible to tackle administrative problems in ways that were not previously possible and lowering costs. It is therefore understandable that, even outside of periods of austerity governments will seek to procure and deploy AI. Examples of current applications include:

- Use of AI driven chatbots to assist patients with chronic conditions in the UK;
- Use of AI in Japan to analyse medical test results and recommend appropriate treatments;
- AI-enabled traffic lights in Pittsburgh;
- The use of analytics in Singapore to reduce traffic congestion;
- The use of chatbots to answer queries from human services case processing officers in Australia;
- The use of machine learning in the Netherlands to detect fraud in social benefit programmes;
- In Chicago the use of spatial algorithms on 911 call data to predict where and when violent crimes are most likely to happen;
- Use of AI to predict how disease will spread over time. Dynamic neural networks were used to predict spread of the Zika virus in 2015 and (at the time of writing) these were being re-trained using Covid-19 data.

These examples are only the tip of a large iceberg. These types of application are attractive at a number of levels, but there are potentially serious problems with the use of AI in many situations. For example, AI currently cannot generally take account of nuances that human agents may detect. Failure by machines to do this may have detrimental outcomes for many vulnerable populations including, for example, those requiring affordable housing, food assistance benefits or children at risk. The Nash report (2017) into the death of a child in Los Angeles found that an autonomous decision system failed to identify the child as being at risk. Instead, reflecting encoded bias, the system focused on poor children from working class communities, especially communities of colour. Similarly, automated essay scoring systems have been shown to encode bias against certain ethnic groups (Eubanks, 2018). These exacerbate existing inequalities in relation to further education and employment opportunities.

2.3. Specific problems with social decision-making algorithms

The literature on the impact of algorithmic decision making on people, organizations and societies is relatively new, but growing rapidly. In recent years, a number of books have explored this phenomenon. One influential contributor to this debate is O'Neil (2016). O'Neil uses the word "model" rather than algorithm. The words do not have the same meaning, but in this context they can be considered as interchangeable. "*Models*", writes O'Neill, "*are opinions embedded in mathematics*" (p. 12). Although her focus is on the private sector, she also argues that the algorithms built into models and the data that they use can create biases and distortions in public policy. For example, she points out how algorithms

can be biased against certain types of college or job applicant. Using other examples O'Neil discusses how machine learning systems that use historical data can learn the wrong things and end up embedding the prejudices, or even the now illegal practices, of an earlier era. A further shortcoming of some machine learning systems is that there are no feedback mechanisms that tell them whether the decisions that they made were good or bad. Even with supervised learning, there is a risk that any biases the supervisor has will be absorbed by the machine. Consider a system which recommends whether or not to hire somebody for a job (such systems exist – see Stanojevic (2020)). If a poor applicant is accepted, then there is at least the possibility of feedback that the decision was not a good one because the employee's performance will be monitored and that performance can in theory (if not necessarily always in practice) be fed back to the software. However should the computer reject a potentially outstanding applicant, the system will never know. As a consequence, it never learns from this type of mistake.

A further problem is that it is often impossible to question or contest the decisions made by certain systems. This particularly the case when, as Whittaker et al. (2018) note, the systems are developed by private vendors and are not subject to public testing, review or auditing to ensure their fitness for a domain. Such systems may be built without explicit public oversight or accountability and are often opaque, a problem that makes discovery of erroneous automated outcomes more difficult. In support of this fact, Whittaker et al. reference the use of a third-party automated decision-making system by Houston Independent School District that used student test data to make decisions regarding teacher employment, promotions and terminations. It transpired that no employee in the School District was able to explain the determinations made by the system despite legal attempts to make them do so. The vendor refused to provide any access to enable experts to determine how the system worked. This type of lack of transparency and explicability, the growing accountability deficit, favours those who create the decision-making technologies at the expense of those who are affected.

Such cases highlight the need for greater transparency, oversight, and accountability in the adoption, development, and implementation of automated decision-making systems. They also highlight the need to reflect on the economic and policy factors that underpin their usage by government, specifically whether their usage is for the benefit of society. Whittaker et al. propose that automated decision systems fairness research needs to be expanded beyond a focus on mathematical parity and statistical fairness towards issues of justice, a view that we strongly endorse. Finally, it is worth noting that this problem is now widely recognised. For example, the EUHIEG Guidelines on Trustworthy AI (EUHLIG 2019) propose seven key requirements that AI needs to meet namely:

1. Human agency and oversight;
2. Technical robustness and safety;
3. Privacy and data governance;
4. Transparency;
5. Diversity, non-discrimination and fairness;
6. Societal and environmental well-being;
7. Accountability.

In considering each of these areas, the question of risk is important.

2.4. Summary

In summary, when it comes to the use of AI in public decision making, significant concerns remain regarding lack of due process, accountability, community engagement and auditing. Whittaker et al. (2018) reference another case where an individual in Arkansas in the United States was assessed for disability

carer supports. The initial nurse-led assessment in 2008 determined that the woman in question required 56 hours of home care help per week. However, in 2016 the automated decision system that was now employed in Arkansas to make these assessments, (using a proprietary algorithm) calculated that the same woman only required 32 hours of carer support per week, a drop in the level of care that had a detrimental impact on her quality of life. By the time use of this algorithmic allocation programme was subsequently successfully subjected to legal challenge and determined to be erroneous and unconstitutional, irreparable damage to the lives of many citizens had already been done.

From the perspective of this paper, a particularly powerful and relevant critique is provided by Eubanks (2018).[3] Eubanks considers a number of cases of public sector computer systems that affect the lives of people including the allocation of housing in Los Angeles and child abuse prediction in Pittsburgh. She shows how computerised decision-making systems are simply too rigid and inflexible for the complex real-world problems with which they are trying to deal. The result is that people are denied treatments, benefits or even access to services to which they are entitled. Wachter-Boettcher (2017) argues that it is all too easy for biases to become "baked into" applications and systems. These include the prejudices of policy makers, but can also include the prejudices of designers and programmers as problems inherited from the data are used to train machine learning systems. Noble (2018) discusses the problem of racial bias built into search engines. While this is a separate issue, the sources of bias she describes are not limited to Google and Bing.

Decision making algorithms have many uses and many advantages. Indeed, such algorithms are at the heart of what has been called *digital-era governance* (Margetts & Dunleavy, 2013) and similar ideas about what Williamson calls *knowing public services* (Williamson, 2014). However, as ever more decisions are delegated to, or influenced by, machines, so grows the need for systems of regulating, controlling and monitoring those machines. Leadbeater (2011, p. 18) describes a system of "*government by algorithm*" that uses "*systems to mine and analyse data, to make automated decisions about allocating resources.*" Demos researchers Wind-Cowie and Lekhi (2012, p. 63) argue that big data:

> "... should be viewed as a transformative agent that has the potential to revitalise, reinvigorate and renew public services."

As Williamson (2014) puts it (p. 18):

> "Algorithms that can be taught to anticipate and predict how people act, or machine learning, have been positioned as a major component of the new techniques of governing promoted by these think tankers and intermediaries."

On the other hand, these government databases do not merely represent the individual. Instead, as they continually join together different pieces of data, they constantly remake the identity of the individual, anticipate likely future behaviours and requirements, and continually recalculate decisions about the best lines of intervention. As goes the private sector (Zubhoff, 2019), so goes the public (Fredridksson et al., 2017; Klievink et al., 2017), although the public sector's use of big data is more likely to be constrained by legal requirements (such as the General Data Protection Regulation in Europe). The individual is a composite enactment made up by software out of data rather than a fixed and stable identity. Williamson (2015, p. 100) draws attention to the point made by Ruppert (2013) that:

[3]There is a talk by Virginia Eubanks available on YouTube in which she elaborates on the ideas in her book at: https://www.youtube.com/watch?v=4xzDgesK2wU.

"... such software systems or 'algorithm machines' do not merely implement a policy or programme but are generative of both their subjects of governing and modes of intervention" (Ruppert 2013: p. 100.).

The literature on AI or algorithmic decision making in complex public sector problems is still embryonic, but a number of clear messages emerge from it. Six important points to emerge from the current literature which are important in the context of this paper are:

1. Faced with complex human or social problems, it is almost impossible for AI decision-making systems to be objective.
2. Feedback mechanisms are often absent or partial, so systems do not learn or alternatively may learn the wrong things because they use historical data which may not be relevant to the current situation.
3. Values need to be embedded in the code of many systems which means that existing human biases and values can also become embedded in the code – sometimes consciously, but also subconsciously.
4. There are significant negative and damaging consequences for people/citizens currently occurring as a result of the use of machine decision-making.
5. There is a large incentive to use such systems to displace more expensive human decision makers.
6. In addition to problems with equity and fairness, there are problems with transparency, explicability and accountability, particularly when such systems are outsourced to the private sector.

To this we would add, as already noted that the risks are not yet always well understood. Addressing these problems is not easy. In the following sections we put forward a framework and some ideas about how to limit the potential damage caused by such systems.

3. A risk assessment framework

We first propose a framework for categorising decision-making algorithms based on two elements:

- A two by two typology;
- A three-dimensional risk spectrum.

Both of these are described in the following sections.

3.1. A two by two typology

All algorithms can first be classified by whether or not they use subjective data. This statement immediately raises the question: what do we mean by 'subjective' data?

All data used by algorithms, whatever its source or whatever form or format it takes, can be classified as objective or subjective. This has parallels with what is generally referred to in the decision support literature as "hard" and "soft" data. As Saaty (2000) puts it hard measurement must be consistent with what is known in nature, whereas soft measurement must be consistent with what is known in the mind. Of course much data depend on scales and categories created by humans, but the distinction we will use is that objective data is not dependent on, or arrived at using human judgement be that of designer programmer or (with some qualifications – see below) user.

An objection to this definition is to claim that all data involves a degree of subjectivity (an argument that has parallels in the longstanding debate about constructivist and positivist research). Almost everything physical that we measure, height, weight, temperature, speed and so on is based on scales or criteria designed by humans and some of these scales are only analogues for more complex phenomena

(temperature being a good example). Nonetheless we argue that there is a tractable difference between such data and personal or social or psychological variables such as happiness, poverty, tolerance, risk aversion and so on. In fact, many of the latter are constructs rather than variables in the formal sense and can only be measured indirectly or by proxy. They are much more complicated to measure and dependent on scales that are typically fuzzy and often arbitrary. We argue that not only is this a valid distinction, but that, for all practical purposes, it is binary.

There is, to take one example, a clear distinction between, say, the airspeed of an aircraft as measured by a sensor and, say, a police officer's assessment, in a bail application, of a suspect's flight risk. This is not just a distinction between something measured by an instrument and something "measured" by a human. Headcounts are objective data even when the counting is done by a person, but a headcount is fundamentally different from data based on an opinion or a judgement, no matter how informed that opinion may be. A simple test is whether presented with the same environment or set of circumstances, a properly calibrated measuring device will always give exactly the same answer. In this context, how the scale is arrived at is unimportant. There is a categorical difference between measuring a patient's body temperature and asking a patient to rank her pain on a scale of one to ten. Most of the data collected by sensors fall into the objective category, but so do other data such as (as already mentioned) counts of various sort or financial data such as cash paid. Subjective data encompasses a wide range of types including constructs, assessments, value judgements, human estimates, opinions and so on. It should be noted that the term "data" is not confined to data that is input by the user when an algorithm is run, it can be, and frequently is, embedded in the algorithm itself.

We acknowledge that concept of objective data is not problem free. A specific problem in AI for example is unstructured data (such as a photograph or video). A photograph can be objective in the sense meant here if it does not directly involve human judgement, but an AI system seeking to make "sense" of a photograph may derive, for example, from supervised learning that may have been be biased to seek out particular features. Of particular relevance in this context is facial recognition.

Secondly, algorithms can be passive in the sense that they solely inform human decision-making or active in the sense that machines make decisions whose consequences are enacted without further human review or involvement or approval. Historically, most decision-making algorithms have been passive. This is what is meant by the term "decision support system" – in itself a large field of both academic research and commercial products (Power, 2008).[4] Algorithms underlie most decision support systems. More recently, advances in technology have resulted in a rapid expansion in the use of active algorithms. Most modern commercial jet aircraft depend on software that both makes decisions and executes them. Active decision-making algorithms have become an important concept in "e-governance" (or digital governance) (Sharma, 2019; Plantera, 2017). As use of algorithms in governance and public administration expands, it is important to understand the implications of ceding the authority to make decisions to machines. What types of decision can be safely outsourced to AI systems and what should be the province of human decision makers? Equally important, how do we exercise control over these machines?

The above analysis leads to the two by two framework illustrated in Fig. 1.

The following are some illustrative examples of each classification. Further examples and their implications are discussed elsewhere below.

[4] A search for "Decision Support Systems" in Google Scholar yields over 750,000 hits.

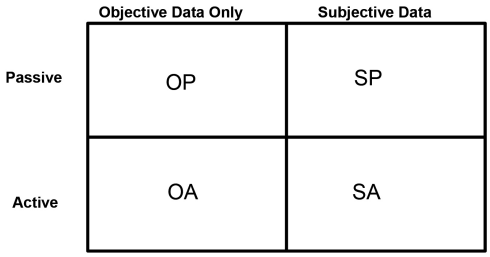

Fig. 1. Classification of algorithms.

3.1.1. Objective/passive (OP)

A simple example of an objective/passive algorithm used in our country is one to compute a citizen's pay related pension entitlement. This is based on a number of factors, primarily on the number and type of "stamps" that a person's various employers have paid for during their working life. There are also technical adjustments that can be made in certain circumstances. The computation is based on purely objective data with no human judgements involved. The rules are explicit. This is a simple example, but public administration is replete with such algorithms. Much of management science/operations research falls into this category e.g. techniques like linear programming or the North-West corner algorithm. While the outputs or results of such algorithms can be implemented automatically, they are typically inputs into human decisions.

3.1.2. Objective/active (OA)

Objective/Active algorithms are widely used in a variety of areas including the operation of machinery. The example of the airspeed sensor given above is just one of numerous sensors on which modern aircraft rely. In normal conditions, most modern commercial aircraft can, for example, land without any input from or action on the part of the pilots. This process depends on numerous algorithms and sensors as well as data transmitted to the aircraft from systems on the runway and air traffic control. Algorithms of this type are found everywhere. A lift/elevator will not move if the weight of the passengers is greater than its certification permits; a police camera will take a photograph only if a vehicle exceeds a certain speed an automatic car will change gear when appropriate and so on.

3.1.3. Subjective/passive (SP)

Subjective passive algorithms are based on human judgements or decisions which may be quite arbitrary or, in some cases, deliberately biased. Consider a Google search. The results of a search request returned by Google are determined by a whole host of factors built into the ranking algorithm – something that has been decided by Google's engineers. When the algorithm is changed, 'before and after' searches may present a different ordering of results or even present different results entirely. Google changes its search engine all the time, 3,234 times in 2018 alone (Moz, 2019). Search algorithms are passive because they pass their output back to the user and how the user uses that information (for example to book a hotel, to buy a product, to find out how to get somewhere, etc.) is up to the user. Most expert systems fall into this category. Recent discussion of medical diagnostics being undertaken by computers has provided evidence that computers are already more accurate than most human doctors in the diagnosis of certain illnesses such as skin cancer (Budd, 2019). So far however, use of such systems requires presenting human doctors with the diagnostic findings and allowing the doctors to take it from there. This will typically take the form of various possible diagnoses with probabilities attached. Basic forms of such expert systems are available on the Web. WebMD is one such example. Such systems are becoming

steadily more sophisticated.[5] One problem with this type of algorithm can be the reluctance of humans to question the system. Human validation of machine decisions is only useful in as much as humans think carefully about what the machine is telling them.

3.1.4. Subjective/active (SA)

While all four types of algorithm are of concern, the fourth category, subjective/active is the most significant from the point of view of risk in public administration. This is where a computer makes a decision based on subjective data supplied by humans and executes a policy decision based on those data. These data often include criteria, thresholds or scores that the system uses to make a decision and to act accordingly. A good example is so-called job pre-screening. When you apply for a job in certain companies today, they will ask you to go online and fill in a form and/or questionnaire. Based on your answers, you may, or may not, be called for an interview. Many such systems today also gather data not supplied by the candidate directly, but extracted from social media and other on-line sources. This may not only be used to rank a candidate, but to decide whether the candidate is a good "fit' for the job (Berkelaar & Buzzanell, 2014). Such algorithms are being used in situations where the consequences are more serious, for example in determining whether a child should be taken into care or the length of a prison sentence. They are also being developed for complex judgement tasks such as examination marking of textual answers (Ab Aziz et al., 2009; Shermis & Burstein, 2013; Nwosu, 2017; RM Results, 2018).

All of the four types have certain risks associated with them though those associated with objective/passive algorithms are probably fairly modest. In section four we examine some of the risks that the other three types, and in particular SA algorithms present in a public sector context. Before doing that, we first consider the risk spectrum.

3.2. A three-dimensional risk metric

In considering the risk(s) presented by an algorithm, three dimensions need to be considered.

The first is *transparency*. Transparency has long been the subject of much discussion (Bannister & Connolly, 2011; Grimmelikhuijsen, 2009). Transparency can refer to data or to process or in the case of algorithms, to both. Meaningful transparency in this context is the ability to see the code/steps in the algorithm so that that people can understand how does its job and how it arrives at its conclusion and/or its decision(s). A critical distinction needs to be made between the ability to see what an algorithm is doing and the ability to understand or explain what it is doing (see below).

Google's page ranking algorithm is a good case in point. What determines the order of the results that you see when you do a Google search? Google does not disclose this although it makes certain information available. Obviously, the main reason Google does not publish details of the algorithm is that the Page Rank algorithm is proprietary and valuable commercial property. Another reason is that it wishes to prevent web site owners gaming the system (a common practice in the past and still tried today). A public system is, or should be, different and yet public bodies have refused, often for what seem like spurious reasons, to publish the code in the algorithms that they use (Carlson, 2017; Whittaker et al., 2018; Eubanks, 2018). It is also important that the code or logic underlying the algorithm be made available in a readable form. The more opaque an algorithm, the higher the risk of unforeseen consequences. Unfortunately some algorithms are black boxes where those affected have no way of

[5] And the bane of physicians lives as more and more patients arrive in surgery convinced that they know what their problem is.

knowing how a decision or recommendation is reached. In extreme cases, the authors of the algorithm may not even know (Veale et al., 2018).[6] It is important to acknowledge that citizens have always sought to game public services, from housing lists to tax loopholes, but this should mean changing the rules to prevent such behaviour, not trying to hide them.

A second criterion is *comprehensibility*. Users (or experts acting on their behalf) should be able to understand *what* an algorithm does and explain how it arrived at its decision. Transparency is of limited value if the code is readable, but indecipherable. There is an important distinction between the algorithms used inside machines, those used in business and those used in the public sector. For example, in the case of avionics or diagnostic machines it is not expected that users (pilots or doctors) will understand the computer code, but they need to understand what the system does and how it behaves in different conditions. To ensure that such systems behave properly and as expected, organisations like the Federal Aviation Authority and the Federal Drugs Administration and their counterparts in other countries certify these systems.[7] Businesses, like Google or Microsoft, may be subject to regulation in some areas, but not in others and are under no obligation to publish the details of the algorithms that they use. A bank may refuse somebody credit based on a proprietary scoring algorithm. In many countries, the bank would not be under any legal obligation to explain to the customer why he has been refused credit. However, when a public sector system makes a decision or recommends one, it should be incumbent on the public body concerned to explain how and why the decision was made. Unfortunately, not only do some public bodies not do this, some of them are unable to do it. Reasons for withholding explanations vary, but tend to include confidentiality and preventing people from gaming the system. Regardless of the rationale, the risk is that in the future a reason for withholding an explanation may turn out to be because the agency simply does not know what the explanation is. It is not just transparency that matters in this case, it is understanding what the code is doing. Ananny and Crawford (2018) refer to this problem as "*seeing without knowing*".

The third criterion is *impact*. The impact or influence of an algorithm on individuals, groups or even societies varies from unimportant to (literally) existential. Negative impacts range from being a source of minor irritation to (again, literally) life-endangering. Table 1 contains is a list of examples of impacts roughly ranked from the marginal to the highly significant.

If one goes to a government website and searches for a particular service or document, then the efficiency and effectiveness of that search does matter, but the order in which search results are presented is rarely of importance. At the other extreme, where algorithms are used to target specific electoral messages to certain voters, there is a potential for major distortions to democracy.

Defining a scale of this nature will require some thought and debate. Similar scales are used in other areas where direct measurement is not meaningful. Two well-known examples are subjective pain scales (Pesudovs & Noble, 2005) or the hunger scale (Derby Community Heath Service 2020). Table 2 is not proposed as a definitive scale, but is indicative of what such a scale might look like.

4. Governing algorithms

Governance has been a topic of growing interest for quite some time and, as is seen all too frequently,

[6] An interesting example is the Irish Tax Authorities who, when they started implementing an entirely new computer system in the 1990s, had to leave one part of the system on legacy machines because nobody understood the code and trying to re-write it was too risky. Eventually, not only did nobody understand the code, nobody still working in the Authority was familiar with the tax laws encoded in the system (though the legislation was still on the books).

[7] It is worth noting that the delegation of this role by some authorities to companies has been much criticised, most recently in the case of Boeing, but reliance on self-regulation was a factor in the lead up to the banking crisis in 2007/8.

Table 1
Range of risk

Impact	Example
Nudging: (at worst irritating)	Targeted advertisements Tailored deals/special offers Biased (often disguised as 'personalized') search results Recommendations (books, hotels, friends) News that is "recommended for you" Rate of interest that you are charged on a loan Whether you get a loan or a credit card Whether you get a job interview Whether you receive a social security payment Whether your health insurance pays out Whether you can get insurance at all What jail sentence you are given Whether your child is taken into care
Potentially life changing (at worst, destructive)	Triggering a global financial crash Undermining democracy

Table 2
Outline possible structure for a risk scale

Title (subcategories)	Description/examples
Nudge (public, personal, mild, strong)	Something designed to encourage an action or behaviour, but aimed at the general public. Public advice. Stronger nudges might contain specific warnings (as for many COVID-19 announcements).
Minor Impact (short, medium, long term)	Decisions that could affect somebody's income in a minor way, for example a rate of interest or the amount of a benefit paid.
Moderate Impact (short, medium, long term)	An example is something that may have a serious long term impact, but not a life threatening one. A contemporary example is the algorithms used to assign grades to students who were not been able to sit state examinations due to COVID-19. A lower grade may mean a university place missed and lifetime changed.
Serious Impact (short, medium, long term)	Decision on whether somebody gets free legal aid or a housing grant. Longer term decisions on what pension a citizen is paid or allocation of public housing.
Life altering (short, medium, long term)	Sentencing systems. Decision on taking children into foster or temporary care. Longer term examples include permanent loss of driving licence or denial of an application for adoption.
Dangerous (individual, society)	An algorithm that has the capability to decide whether to give a death sentence or who gets medical treatment or is first to get a new vaccine. Algorithms that could endanger democracy or individual liberties (in, for example, e-voting systems).

where governance is poor or absent it almost invariably results in undesirable consequences. The use of decision-making algorithms (and AI generally) requires proper governance and this topic has recently started to attract academic attention in the legal and philosophical literature (Coglianese & Lehr, 2018; Lepri et al., 2018; Fink, 2018; Valentine, 2019).

There are three components necessary for such governance:

1. Risk assessment and management
2. Accountability and
3. Regulation.

We discuss each of these in turn.

4.1. Risk assessment and management – Using the framework

The framework whose elements have been described in the preceding section can be used to assess the risk associated with a given algorithm or AI system. Risk is assessed under four headings:

- Type of algorithm;
- Transparency;
- Comprehensibility; and
- Impact.

The type of algorithm affects risk. OP algorithms, while not always risk free, are generally low risk. Such algorithms can be faulty like any other, but subjective bias will only occur when the user's own internal bias(es) affects a decision on the basis of what the algorithm has told the user, though there is an increased risk when a bias in the machine coincides with a user's own bias.

OA algorithms are somewhat higher risk, but the risk lies in possible design and programming flaws rather than in built in bias.

SP algorithms are higher risk again as they depend both on the risk of internal bias within the machine and on the human user's ability to use what the machine tells him or her effectively and to use it to inform, but not override his or her own judgement. A system recommendation of a sentence for somebody found guilty can be modified by a judge for example. The difficulty is that many humans may be reluctant to override a machine recommendation. One reason for this may be legal or personal exposure; it can be safer to hide behind the algorithm. An analogous problem is over-testing in medicine. Some doctors order large numbers of tests, not because they are necessary for diagnosis or action, but because of the fear that if something goes wrong, they will be sued for having failed to explore every possibility. By relying on the computer, responsibility can be transferred to the computer and by implication to the system designers and coders. Where, as seen in the literature review, the system provider is a private company that refuses to disclose details of how its decision-making algorithms work, then the responsibility falls into a black hole – similar to the excuse that is often used in the public sector of 'systemic failure' where the avoidance of individual blame is key.

SA algorithms present the greatest risk. In such systems, the decision is entirely abdicated to the machine.

It would not be difficult to develop a scoring system based on these, but it would be better if expert actors looked at a given decision-making system under each of these headings and assessed risk on a spectrum. A black box system based on subjective scoring which can affect people's lives in fundamental ways would be high risk; a support system which offers powerful 'what if' type exploration of the potential consequences of different decisions is likely to be lower risk (though not risk free).

One does not have to go far to find systems that, by any standards, are medium to high risk. One case which has been in the news recently is the UK Home Office's system for processing visa applications (Warrell, 2019). This is a SP type system; the algorithm streams and ranks applicants into red, amber and green according to their level of perceived risk and then passes this information to humans to decide whether or not to approve the application. The Home Office has refused to disclose details of which factors are used to assess this risk or how they are weighted. Such an algorithm is ineluctably based on a large amount of subjective judgement as, of its nature, statistical data on such risks will be incomplete. Commenting on the system, Christine Blacklaws, president of the British Law Society observed that such a system:

> "... may well disadvantage certain groups of people based on generic markers such as age, country of origin or whether they have travelled before".

But the public does not know. This raises the question; should the public know? There are circumstances when transparency is clearly not in the public interest (Moon & Adams, 2015). In this case, transparency might enable people to game the system. This possibility therefore has to be factored into any risk assessment.

In other words, just because an algorithm is an impenetrable black-box or has its code available on-line does not, of itself, make the system high or low risk. It depends on the purpose of the algorithm – if an error in such a system does not have major consequences the risk might be ranked low. Nevertheless, while risk is determined by a combination of factors, use of opaque algorithms is, with rare exceptions, not a good policy.

4.2. Accountability

Algorithms require policing (Veale et al., 2018) and that means that there needs to be an accountability framework. Accountability can be a problematic concept at the best of times (Dubnick, 2002). It is based on the principle that individuals and/or organisations are answerable for their actions and that failure or bad behaviour will be identified, exposed, where possible corrected and, where appropriate, punished. In practice, organizations frequently seek to hide problems, are reluctant to expose themselves to full scrutiny even when problems are identified and are adept at avoiding punishments that should be imposed on those responsible. True accountability is often lacking in the public sector (Guerin, 2018) and even where it exists, sanctioning of the individuals responsible for failure frequently does not happen. Politicians often suffer as a result of administrative failures over which they have little or no control or influence while those actually culpable for such failures hide behind the shield of public service anonymity

Ignoring for the moment the many real-world shortcomings of accountability in practice, accountability means, in principle at least, that some person or persons is answerable for a decision and if that decision is bad then sanctions can be applied to a human being. This may range from public criticism to demotion, firing, a fine or, in extreme cases, a jail sentence (or, in the case of politicians, having to resign or losing their seat at the next election).

In any bureaucracy, the standard way to avoid personal sanction is to blame the system. The term 'systemic failure' has already been mentioned and is a commonplace in public life where numerous failures of public policy or practice have successfully been blamed on the 'the system' by individual public servants and managers (Anderson, 2009; O'Brien, 2012; Leon & Orriols, 2019). In some countries, it is common for government ministers or senior managers to resign and take the blame for failure on the part of their subordinates. In others it is rare, although it occasionally does happen (Bardon, 2017).

But what happens when the failure or scandal is caused by an algorithm? In such circumstances, the term 'systemic failure' acquires new possibilities. Who should carry the can? Consider the Boeing 800 Max disasters. It was a software failure (or a combination of software and hardware failures) that led to the two tragedies. If we want to attribute blame for this to a human (or humans), who should that be? There is a long list of candidates including:

- The Federal Aviation Authority (who delegated testing to Boeing);
- Boeing senior management (for deciding to rely on software to fix an inherently unstable configuration rather than re-designing the aircraft);
- The aircraft's designers (who, inter alia, specified only one speed sensor creating a single point of failure);
- The managers who decided not to have a second sensor in order to reduce costs;
- The software designers;

- The programmers who wrote the code;
- The project managers who failed to test the system adequately;
- The people in Boeing who failed to put in place proper training for the pilots.

There are undoubtedly others. The only people who were not responsible were the pilots (who of course were the first to be blamed by Boeing).[8]

It is not hard to translate this catalogue of possible people responsible into any one of a number of public service contexts.

One way of dealing with this would be to define a clear structure of what might be called *responsibility points* for any public sector decision-making system. At the moment, the single point of accountability is usually a politician (or a carefully chosen scapegoat – often well down the line). Responsibility points would assign both responsibility and accountability and clarify the borders/parameters of that responsibility for everyone concerned with a decision-making system and with the decisions it makes. A possible structure might include:

- Those who determine policy;
- Those who specify the requirements to implement the policy;
- Project owners;
- Project managers;
- Designers;
- Coders/programmers;
- Testers;
- Managers/administrators;
- Final Decision makers;[9]
- Front line staff.

A critical point is that there should be a person or persons responsible or answerable for every decision and for every stage in the process that leads to the final decision. Furthermore, it should never be the case where blame for failure cannot be pinned on one or more people, i.e. there must be no gaps in the accountability chain. Where a human makes a decision, whether based on machine 'advice' or not, the decision maker is identifiable. However where a machine makes the decision, then if a decision turns out to be wrong or harmful, there should be an clear audit trail back to the person or persons responsible for this whether it be the politician who signed off on the policy or the programmer who wrote the code.

It is essential that it be clear from the outset where responsibility lies. An analyst or programmer cannot be expected to be aware of potential nuances in a wicked social problem or even a relatively simple problem like visa approval. The responsibility of the programmer is to code whatever the specification specifies. Furthermore, it needs to be clear that there are real sanctions for errors that are appropriate to the nature of the error and its cause as well as allowing for the degree of discretion or authority that the person in question has. For example, if a system recommends a decision to a street level bureaucrat, can that bureaucrat reject the advice or seek more information or does she have no choice but to implement what the machine tells her? In addition, there need to be processes in place where humans can query and refer such decisions upward without fear of reprimand.

Finally, sanctions need to be appropriate up to and including criminal indictment in extreme cases.

[8]It is worth noting that, at the time or writing, nobody in Boeing has resigned as a result of these disasters. One senior executive has 'retired' (https://onemileatatime.com/737-max-executive-retires/).

[9]There will, of course, be decisions made by people in all of these groups. People should be responsible for the decisions that they make, but in a proper system, to adapt the words of Harry Truman, the buck has to stop with one person or maybe group.

The principles outlined are pragmatic and, if done properly, effective. Given human nature, they are also likely, like many attempts to introduce good governance, to encounter resistance, but they are essential if there is to be public trust in such systems and processes.

4.3. Regulation

Thirdly we propose that there should be a regulatory agency responsible for approving all algorithms which affect people's lives in a material way. This is simply the extension of a principle and praxis that applies in many other areas of life. No medical drug may be prescribed or sold without the approval of the appropriate authorities (e.g. in the USA, the Federal Drugs Agency, in the EU, the European Regulatory System for Medicines). The same is true for all sorts of other products from aircraft to zoos. Often such regulation is multi-layered. Drug regulators not only ensure that new drugs have been extensively tested, they also supervise their production and increasingly today the supply chain that delivers them to avoid forged or fake drugs entering the system.

What is needed is an agency to monitor and approve the use of algorithms, AI or not, in routine public decision making, say an Artificial Intelligence/Algorithm Regulatory Agency (AIRA). The remit of such an agency would be decision-making algorithms which have the potential to materially affect citizens, business or other organisations on a continuing basis. It would not include, for example, decision support systems or models built for one-off decisions. Such an agency could certify algorithms before they are used, certifying that:

- They have designed and developed according to established software engineering standards;
- They are properly documented;
- They are appropriately transparent;
- An appropriate responsibility framework is in place;
- They have been thoroughly tested;
- Proper training courses are in place.

An AIRA would also have an appeal function whereby citizens, unhappy with a decision or the explanation for that decision or in the case of a dispute, could appeal a decision to the agency for arbitration.

4.4. Other possibilities

In addition to risk assessment, accountability and regulation, there are other actions that should be considered. One is to mandate that all public sector decision-making software other than that related to areas of state security and policing, should be open source. Another possibility is a charter of rights with regard to machine decision-making. How does a citizen appeal against a decision made by a machine? These, and other possibilities, require further exploration.

In some ways, the problems presented by contemporary AI and algorithm-based decision-making are new; in others they are not. There have always been problems with (non-AI) algorithms in "traditional" ICT systems; the history of e-government is littered with computer generated problems (and disasters). However, a key difference with AI is that it "learns" and this creates the possibility that it will learn the wrong things and/or that it will amplify errors in the original design. Existing faulty algorithms repeat the same mistakes. AI algorithms may have the capability to make new ones. AI and the drive to reduce costs and increasing efficiency in the public sector is going to make the use of this technology increasingly pervasive in the public sector and whatever the provenance of the problems that are ineluctably going to ensue, it is preferable that they should be anticipated rather than reacted to in retrospect.

5. Conclusion

Artificial intelligence (AI) takes many forms and it is not difficult to think of developments in AI that have or will have implications, and often major implications, for public administration, public governance and/or democracy as well as for individual citizens.

One manifestation of AI that is already generating extensive debate is the increasing use of algorithmic decision-making for implementing public policy in a wide range of areas including, for example, medicine, social work education justice and public housing. A growing number of scholars and commentators have expressed concerns not just about machines making bad decisions resulting from poor design or coding, but about more serious longterm risks such as inbuilt bias and the permanent embedding of existing social prejudices and power structures in software. Decisions may also be influenced by the subconscious prejudices of the individuals who formulate, clean and structure the data that is used in such systems.

At another level, there are questions about the ability of such systems to adjust to unforeseen circumstances or to be able to exercise the types of discretion that human actors often do when faced with complex and nuanced social situations. The absence of human factors such as sympathy, empathy, insight, experience and flexible thinking can lead to rigidity and potentially harmful consequences as machines work with hard coded (and what may be poorly understood) rules as well numeric ranking systems that try to reduce complex decisions to simple weights or scores.

Being able to detect and categorise these risks is essential in order to develop a systematic and calibrated response. In this paper we have proposed a risk management framework comprising a two-by-two typology and a three-dimensional risk spectrum that is designed to assist scholars and practitioners interested in calibrating and managing such risks in the absence of structured accountability and legal regulation of AI in the public sphere. Society needs to move fast on this. The speed of development of AI is accelerating and there may come a tipping point whereby control of AI becomes impossible. We are, just about, still at a stage where this can be controlled. We need to act now.

References

Ab Aziz, M. J., Dato'Ahmad, F., Ghani, A. A. A., & Mahmod, R. (2009 October). Automated marking system for short answer examination (AMS-SAE). In *2009 IEEE Symposium on Industrial Electronics & Applications* IEEE, Vol. 1, 47-51.

Acemoglu, D., & Restrepo, P. (2017). Robots and Jobs: Evidence from US Labor Markets, NBER Working Paper No. 23285. March 2017. Available at https://www.nber.org/papers/w23285.

Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, *20*(3), 973-989.

Anderson, J. (2009). Illusions of Accountability: Credit and Blame Sensemaking in Public Administration. *Administrative Theory & Praxis*, *31*(3), 322-339. Available at: http://www.jstor.org/stable/25611001.

Arntz, M., Gregory, T., & Zierahn, U. (2016), "The Risk of Automation for Jobs in OECD Countries: A Comparative Analysis", *OECD Social, Employment and Migration Working Papers*, No. 189, OECD Publishing, Paris.

EU High-Level Expert Group on Artificial Intelligence (2019) "*Ethics Guidelines for Trustworthy AI"*. European Commission, Brussels, Available at: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

Bannister, F., & Connolly, R. (2011). The Trouble with Transparency: A Critical View of Openness in e- Government, *Policy and Internet*, *3*(1), 1-30.

Bardon, S. (2107). Frances Fitzgerald resigns in 'national interest' to avoid an election, The Irish Times, 28 November 2017.

Berkelaar, B. L., & Buzzanell, P. M. (2014). Cybervetting, person – environment fit, and personnel selection: Employers' surveillance and sensemaking of job applicants' online information. *Journal of Applied Communication Research*, *42*(4), 456-476.

Bostrom, N. (2014). Superintelligence. Oxford UK: Oxford University Press.

Budd, K. (2019). Will artificial intelligence replace doctors? AAMC New, 9 July 2019, The Association of American Medical Colleges. Available at: https://www.aamc.org/news-insights/will-artificial-intelligence-replace-doctors.

Carlson, A. M. (2017). The Need for Transparency in the Age of Predictive Sentencing Algorithms. *Iowa L. Rev*, *103*, 303-219. Available at: https://ilr.law.uiowa.edu/assets/Uploads/ILR-103-1-Carlson.pdf.

Coglianese, C., & Lehr, D. (2018). Transparency and algorithmic governance. Administrative Law Review, *71*, 1.

Creemers, R. (2018). China's Social Credit System: An Evolving Practice of Control. Available at SSRN https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3175792.

Cunningham, S. M. (1967). The Major Dimensions of Perceived risk in Cox. D.F. (Ed.) Risk Taking and Information Handling in Consumer Behaviour, Boston, Harvard University Press.

Derby Community Heath Service (2020) The Hunger Scale, NHS Trust. Available at: http://www.dchs.nhs.uk/assets/public/dchs/llb/tools/tools_1-11/4_DCHS_A5_4pp_The_Hunger_Scale.pdf.

Dubnick, M. J. (2002). Seeking salvation for accountability. In Proceedings of the Annual meeting of the American Political Science Association, August 29–Setpember 1 2002, Boston. 29, 7-9.

Eggers, W. D., & Beyer, T. (2019). AI-augmented government: Climbing the AI maturity curve. Deloitte Insights. Available at: https://www2.deloitte.com/insights/us/en/industry/public-sector/government-trends/2020/ai-augmented-government.html.

Eubanks, V. (2018). Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor New York: St. Martin's Press.

EU High-Level Expert Group on Artificial Intelligence (2019) "Ethics Guidelines for Trustworthy AI. A Definition of AI: Main Capabilities and Disciplines". European Commission, Brussels. Available at: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

Fink, K. (2018). Opening the government's black boxes: freedom of information and algorithmic accountability. *Information, Communication & Society*, *21*(10), 1453-1471.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Lütge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People – An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines*, *28*(4), 689-707. Available at: https://link.springer.com/article/10.1007%252Fs11023-018-9482-5.

Fredriksson, C., Mubarak, F., Tuohimaa, M., & Zhan, M. (2017). Big data in the public sector: A systematic literature review. *Scandinavian Journal of Public Administration*, *21*(3), 39-62.

Guerin, B. (2018). Accountability in Modern Government, The Institute of Government. Available at: https://www.instituteFloridixgovernment.org.uk/sites/default/files/publications/IfG%20accountability%20discussion%20paper%20april%202018.pdf.

Grimmelikhuijsen, S. (2009). Do transparent government agencies strengthen trust? *Information Polity*, *14*, 173-186.

Karoly, L. A., & Panis, C. (2004). The 21st Century at Work: Forces Shaping the Future Workforce and Workplace in the United States. Santa Monica, CA: RAND Corporation.

Klievink, B., Romijn, B. J., Cunningham, S., & de Bruijn, H. (2017). Big data in the public sector: Uncertainties and readiness. *Information Systems Frontiers*, *19*(2), 267-283.

Leadbeater, C. (2011). The Civic Long Tail, London: Demos.

Leon, S., & Orrios, L. (2019). 'It's Westminster's fault'. Political identities and blame attribution in devolved systems, London School of Economics. Available at: https://blogs.lse.ac.uk/politicsandpolicy/blame-attribution-in-devolved-systems/.

Lepri, B., Oliver, N., Letouzé, E., Pentland, A., & Vinck, P. (2018). Fair, transparent, and accountable algorithmic decision-making processes. *Philosophy & Technology*, *31*(4), 611-627.

Levi, P. (1966). Storie Naturale, Einaudi, Italy (published under the pen name Damiano Malabiaia).

Liang, F., Das, V., Kostyuk, N., & Hussain, M. M. (2018). Constructing a Data-Driven Society: China's Social Credit System as a State Surveillance Infrastructure. *Policy & Internet*, *10*(4), 415-453.

Manyika, J., & Bughin, J. (2018). The promise and challenge of the age of artificial intelligence, McKinsey Global Institute Executive Briefing, October 2019. Available at: https://www.mckinsey.com/featured-insights/artificial-intelligence/the-promise-and-challenge-of-the-age-of-artificial-intelligence.

Margetts, H., & Dunleavy, P. (2013). The second wave of digital-era governance: a quasi-paradigm for government on the Web. *Philosophical Transactions of the Royal Society A*, *371*: 20120382. Available at: http://dx.doi.org/10.1098/rsta.2012.0382.

Mitchell, A., & Diamond, L. (2018). China's Surveillance State Should Scare Everyone, The Atlantic, 2 February 2018. Available at: https://www.theatlantic.com/international/archive/2018/02/china-surveillance/552203/.

Moon, D., & Adams, C. (2015). Too much of a good thing? Balancing transparency and government effectiveness in FOI public interest decision making. *AIAL Forum*, *82*, 28-39.

Moz (2019). Google Algorithm Update history. Available at https://moz.com/google-algorithm-change.

Myers West, S., Whittaker, M., & Crawford, K. (2019). Discriminating Systems: Gender, Race, and Power in AI The AI Now Research Institute, Retrieved from: https://ainowinstitute.org/sdiscriminatingsystems.html.

Nash, M. (2017). "Examination of Using Structured Decision Making and Predictive Analytics in Assessing Safety and Risk in Child Welfare" (Los Angeles: County of Los Angeles Office of Child Protection, May 4, 2017) Available at: http://file.lacounty.gov/SDSInter/bos/bc/1023048_05.04.17OCPReportonRiskAssessmentTools_SDMandPredictiveAnalytics_.pdf.

Noble, S. U. (2018). Algorithms of Oppression: How Search Engines Reinforce Racism. NYU Press.

Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M. E., & Kompatsiaris, I. (2020). Bias in data-driven artificial

intelligence systems – An introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *10*(3), e1356. Accessed 5 August 2020 at: https://onlinelibrary.wiley.com/doi/pdf/10.1002/widm.1356.
Nwosu, J. C. (2017). Comparative Assessment of E-exam and E-marking Integration in selected Universities in Ogun State: Undergraduate Student's Perspectives. *Development and Innovation* Available at: http://www.incedi.org/wp-content/uploads/2018/06/290.pdf.
O'Brien, C. (2012) "No one is responsible for anything", The Irish Times, 5 March 2012.
O'Neil, C. (2016). Weapons of Math Destruction, Broadway Books.
Plantera, F. (2017). Artificial Intelligence is the next step for e-governance in Estonia" State adviser reveals. Available at: e-estonia.com/artificial-intelligence-is-the-next-step-for-e-governance-state-adviser-reveals/.
Power, D. J. (2008). Decision support systems: a historical overview. In Handbook on decision support systems 1 (pp. 121-140). Springer, Berlin, Heidelberg.
Pesudovs, K., & Noble, B. A. (2005). Improving subjective scaling of pain using Rasch analysis. *The Journal of Pain*, *6*(9), 630-636.
RM Results (2018) The power of Artificial Intelligence (AI) in e-marking. Available at: https://rmresults.com/blog/the-power-of-artificial-intelligence-ai-in-e-marking.
Ruppert, E. (2013). "Not Just Another Database: The Transactions that Enact Young Offenders." Computational Culture, issue 3. Accessed February 11, 2014. Available at: http://computationalculture.net/article/not-just-another-database-the-transactions-that-enact-young-offenders.
Saaty, T. (2000). Fundamentals of decision making and priority theory with the analytic hierarchy process (Vol. 6). RWS Publications.
Sharma, T. (2019). How artificial intelligence can improve e-governance services? Insight and Resources: Global Tech Council. Available at: globaltechcouncil.org/artificial-intelligence/how-artificial-intelligence-can-improve-e-governance-services/.
Shermis, M. D., & Burstein, J. (Eds.). (2013). Handbook of automated essay evaluation: Current applications and new directions. Routledge.
Singer, N. (2019). Amazon is pushing facial technology that a study says could be biased, The New York Times, Retrieved from: https://www.nytimes.com/2019/01/24/technology/amazon-facial-technology-study.html.
Stanojevic, M. (2020). Best automated hiring software 2020, Windowreport.com. Available at: https://windowsreport.com/automated-hiring-software/.
Valentine, S. (2019). Impoverished Algorithms: Misguided Governments, Flawed Technologies, and Social Control. *Fordham Urb. LJ*, *46*, 364.
Veale, M., Van Kleek, M., & Binns, R. (2018, April). Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. In Proceedings of the 2018 chi conference on human factors in computing systems (p. 440). ACM.
Wachter-Boettcher, S. (2017). Technically wrong: sexist apps, biased algorithms, and other threats of toxic tech. WW Norton & Company.
Warrell, H. (2019). Home Office under fire for using secretive visa algorithm, Financial Times, June 9, 2019. Available at: https://www.ft.com/content/0206dd56-87b0-11e9-a028-86cea8523dc2.
Whittaker, M., Crawford, K., Dobbe, R., Fied, G., Kaziunas, E., Mathur, V., Myers West, S., Richardson, R., Schultz, J., & Schwartz, O. (2018). AI Now Report 2018 AI Report Institute. Available at: https://ainowinstitute.org/AI_Now_ 2018_Report.pdf.
Williamson, B. (2014). Knowing public services: Cross-sector intermediaries and algorithmic governance in public sector reform. *Public Policy and Administration*, *29*(4), 292-312.
Wind-Cowie, M., & Lekhi, R. (2012). The Data Dividend. Demos. Available at: http://www.demos.co.uk/files/The_Data_Dividend_-_web.pdf.
Zuboff, S. (2019). The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power, Profile Books.