Why Predictive Algorithms are So Risky for Public Sector Bodies

Madeleine Waller[1] and Paul Waller[23]

October 2020

London

[1] Kings College London, Department of Informatics.
[2] University of Bradford, Faculty of Management, Law & Social Sciences.

Abstract

This paper collates multidisciplinary perspectives on the use of predictive analytics in government services. It moves away from the hyped narratives of “AI” or “digital”, and the broad usage of the notion of “ethics”, to focus on highlighting the possible risks of the use of prediction algorithms in public administration. Guidelines for AI use in public bodies are currently available, however there is little evidence these are being followed or that they are being written into new mandatory regulations. The use of algorithms is not just an issue of whether they are fair and safe to use, but whether they abide with the law and whether they actually work. Particularly in public services, there are many things to consider before implementing predictive analytics algorithms, as flawed use in this context can lead to harmful consequences for citizens, individually and collectively, and public sector workers. All stages of the implementation process of algorithms are discussed, from the specification of the problem and model design through to the context of their use and the outcomes. Evidence is drawn from case studies of use in child welfare services, the US Justice System and UK public examination grading in 2020. The paper argues that the risks and drawbacks of such technological approaches need to be more comprehensively understood, and testing done in the operational setting, before implementing them. The paper concludes that while algorithms may be useful in some contexts and help to solve problems, it seems those relating to predicting real life have a long way to go to being safe and trusted for use. As “ethics” are located in time, place and social norms, the authors suggest that in the context of public administration, laws on human rights, statutory administrative functions, and data protection — all within the principles of the rule of law — provide the basis for appraising the use of algorithms, with maladministration being the primary concern rather than a breach of “ethics”.

Keywords

Algorithms, Predictive Analytics, Data Analytics, Artificial Intelligence, AI, Machine Learning, Automated Decision Making, Government Services, Public Sector, Public Administration, Ethics, Trust, Data Protection, Law, Accuracy, Bias, Transparency, Explainability, Accountability

Citation

M. Waller and P. Waller, “Why Predictive Algorithms are So Risky for Public Sector Bodies”, London, October 2020.

Introduction

As technology progresses, there has been an expanding interest in trying to make the operations of public sector bodies more efficient and reliable by applying algorithmic methods, most recently so-called artificial intelligence (AI) and machine learning techniques [1], [2]. These have been used for a range of prediction, classification and data analysis applications. However, problems have arisen in high risk cases (as defined in [3]) covering criminal justice, rail franchising, visa applications, child welfare, and school examination results [3]–[7]. Consequently, much work has been performed based on the deductions that regulations and guidelines for such technologies are falling behind their implementation, and often the stakeholders are not fully involved or do not fully understand the processes being used.

Many public sector and research organisations have produced analyses and guidelines on "Data Ethics" or "AI Ethics" [8]–[10], for example [11], [12], but so far these show little sign of having an impact [13]–[17]. Rather than do the same, this paper aims to move away from the labels of 'Ethics', 'AI' and 'machine learning', often embedded in the "digital" mantra, to bring together ideas and examples across different disciplines to set out the risks and issues in using predictive or data analytic algorithms in government services. This will be done by retracing the life cycle of an algorithm, from how it is used in real life to the original idea to build a model, in order to explore the challenges at each stage.

Bias and data of indeterminate quality are well-established risks; however these are not the only factors to consider (although they are still significant). An algorithm is a set of mathematical and statistical procedures applied to data: it needs skilled design and testing and thus there is a multitude of opportunities to build in errors at all stages of its development and use. Policy objectives, data encoding, accuracy, the nature of the application domain, interpretation of results by professionals, the political and social impact are just some of the potential issues. Existing research into the use of algorithms in children's welfare [18], [19] and exam grading [20] have provided useful case studies of the aforementioned issues of goal setting, data, modelling, and accuracy.

Notably, the accuracy of algorithms in predicting life outcomes has been strongly challenged. What Works for Children's Social Care (WWCSC) analysed case notes on tens of thousands of children, and created algorithms to make predictions about their future [18]. The models missed most of the children at risk, and when they flagged a child as at risk, they were only right 40% of the time. These results are consistent with an even larger study reported in America [19] where 160 teams built predictive models for six life outcomes. The data used covered thousands of items collected on the lives of families, but none of the teams was able to make accurate predictions for cases held back from training the systems.

The extensive existing literature relating to "ethics" has helpfully highlighted several areas of concern, notably the importance of transparency about how an algorithm is applied. One proposed way to examine whether an algorithm is fit for use is centred on considering five factors: beneficence (will it do good), non-maleficence (will it not do harm), autonomy (do humans still have agency), justice (is it legal, fair and honest) and explainability (can any stakeholder understand what is going on and why) [21]–[23]. These can be compressed into the pivotal questions of whether the algorithm itself can be trusted: whether the outcomes are reliable and whether the outcome can be explained to any stakeholder. Determining this can be complex [24], [25], and this paper aims to assist in this matter by identifying the key issues that need to be investigated.

The special nature of Public Administration

Public administration — the function of public bodies or governmental organisations —is a distinct domain of application of algorithms, as public bodies are created and defined by constitutional and administrative law. They perform statutory functions relating to obligations (e.g. paying tax), entitlements (e.g. benefits, healthcare) and duties of care to vulnerable groups (e.g. child welfare).

Algorithms in public administration are a component of an administrative instrument [26] that executes part of the public body's statutory functions. Its legal basis, governance, accountability and transparency requirements are therefore defined by the legal and administrative framework covering that administrative instrument. Any decision by the public body supported by an algorithm is no different in these regards to a decision made by any other means.

In regard to discretionary decisions made as part of such statutory duties, citizens are entitled (under the principle of the rule of law[4] and human rights legislation) to contest and seek remedy for any decision that adversely affects them [27]. Where the impact of algorithm-based decisions makes a difference to people's lives and wellbeing, it is crucial to be able to fully trust the algorithm. The effects of wrong decisions being made can be severe, as demonstrated in the field of criminal justice and child welfare services [7], [28]–[30]. Much of the focus of discussion so far has been on the effects on individuals, but harm can be caused to people collectively as well, e.g. minority groups [31].

Public officials and politicians are accountable, and subject to scrutiny by the judiciary. Any decision-making process that impacts an individual's life and is based on data relating to people is hard to justify without an extensive guarantee that it has been thought through thoroughly. Many public administrative decisions fall into this category.

> *"…determining your future based on someone else's past has much greater implications. The kind of social data that is involved in these critical life decisions is inherently unpredictable."* [32]

Remarkably, Cicero foresaw the 2020 UK exam grading fiasco[5] when he wrote "summum ius summa iniuria" ([33] cited in [27]) meaning "an acritical [mechanical] application of law, without understanding and respect of laws' purposes and without considering the overall circumstances, is often a means of supreme injustice"[6]. In other words, an algorithmic result adopted without taking account of the human context can lead to unjust administrative decisions.

The understanding of "ethical behaviour" depends on social context: time, place and social norms. Hence we suggest that in the context of public administration, laws on human rights, statutory

---

[4] Public bodies, officials and politicians may only act where the law permits them to, otherwise they may be challenged in court. The rule of law provides the common principles of transparency, accountability, predictability, consistency, and equality before the law [49], [69].

[5] Due to the Covid-19 pandemic, no students in the UK took their summer GCSE or A-Level exams. Instead their grades were determined by an algorithm. According to the BBC, teachers supplied an estimated grade and a ranking for each student for every subject. The algorithm then calculated their grades based on these factors and the school's performance in each subject for the last 3 years to achieve an overall pattern of grades similar to previous years' distributions. 36% of student's A-Level grades in England were downgraded from the teacher's predicted one and this negatively affected students at state schools more than at private schools [3], [20], [70].

[6] https://www.latin-is-simple.com/en/vocabulary/phrase/1880/

administrative functions, and data protection provide the basis for appraising the use of algorithms: maladministration is the primary concern rather than a breach of "ethics". Nevertheless, the five ethical factors mentioned above still provide a helpful framework of challenges.

> *"The reasons for a decision must be **intelligible** and they must be **adequate**. They must enable the reader to understand **why** the matter was decided as it was and **what conclusions** were reached on the 'principal important controversial issues'. ... The reasoning must not give rise to a **substantial doubt** as to whether the decision-maker erred in law, for example by misunderstanding some relevant policy or some other important matter or **by failing to reach a rational decision on relevant grounds**."* [30]

Impact on public administrators' professional behaviour

Whilst algorithmic calculations are proclaimed to be helpful to public officials faced with judgment-based decisions (for example in social services or criminal justice), they may in fact pose awkward problems [34], [35]. First, "automation bias" may be present – the tendency for humans to give greater credence to the outputs of technical systems than their own judgement [7], or the opposite — "algorithm aversion" [36]. Second, due to the political and financial capital investment in such a system, it is likely that its introduction will influence decisions more than simply by its provision of individual predictions [37]. The assessment of the cost of an error will change.

For example, if a welfare case is predicted to be one where there is a risk of harm occurring, it is likely that politicians and professionals would be strongly criticised if they did not intervene and harm occurred "when the system said it would". It would inevitably be cast as a professional failure. But numerous no-risk cases can be flagged as at risk by such a system [18], so depending on the system calibration there could be a tendency toward significant unnecessary interventions, with the associated disbenefits.

However, should harm occur in a case not flagged as at risk by the system, then this can be cast as "system error", deflecting blame from professionals. In such a situation however, if it were known that professionals did consider the case to be at risk, but allowed that judgement to be overturned by the system prediction (perhaps because all resources were directed to flagged cases), then matters get complicated and difficult for all concerned. Such situations can arise, and the human tendency may be to take the path of least potential personal implications, leading to a diminution over time of professional performance and critical capability [17], [38].

Many issues that occur when an algorithm interacts with human reality can stem from the very first step — determining the policy or goal specification for the task. The algorithmic calculations and the data provided give the result they are programmed to deliver, which may not be the ones desired. In the exam grading example, the policy goal (a grade distribution similar to previous years) was set in a way that inevitably would lead to chaotic individual results even if (in fact, because) the algorithm was constructed and the program written accurately to meet it [20]. A policy goal is a political decision and should be handled as such by appropriate processes.

"Once the goals were identified, and in common with any other policy, they would need to go through democratic scrutiny, debate and accountability before being implemented." [20]

Engagement of stakeholders from the outset of system design could help to keep things on track, but even this is not fool proof given the potential for unforeseen consequences, and the complexity of the exercise as this paper now discusses.

Accuracy

Despite claims and a human tendency toward optimism, no predictive system of the type considered here is entirely accurate [39]. Taking a "simple" example, the accuracy measures of predictive tests that produce positive or negative (yes/no) outputs are expressed as *sensitivity* (sometimes *recall*, the probability that a test will correctly identify things it should); *specificity* (or *precision,* the probability that a test will reject things it should reject – to not be fooled into an incorrect identification); and *incidence* (proportion of the population of cases that warrant positive identification). Depending on the system, there may be a trade-off between sensitivity and specificity [18], [40]. *Accuracy* may be quoted as a single probability of getting the answer right.

There is strong possibility that there are few practical ways of telling how accurate such systems will be when applied in a real and unique local setting. This is especially true if the system is bought "off-the-shelf" and calibrated from data from other localities. If calibrated on local data, or developed in-house, there would need to be high volume and quality of data available to determine levels of accuracy with any confidence. Accuracy figures, usually given as percentages, are actually conditional probabilities [41]. Many people struggle to work with probabilities, and conditional ones can be very tricky to grasp.

While the accuracy measures for a test are constant for that test, the interpretation of a result from the test for an individual case differs depending on the incidence of positive cases in the population tested [40]. This is not particularly intuitive. It also is quite hard to interpret the meaning of a negative test, as that does not necessarily mean a case is actually negative (unless the specificity is 100%). If the system is run on a population with a high incidence of actual positives, there is an increased chance of a case identified as negative being really a positive one compared to the same test on a population with low incidence. On the other hand, with low incidence, the chance of a case identified positive being actually negative increases rapidly if the specificity falls, even to 80% [42].

Most algorithmic systems produce uncertain results i.e. there is a probability associated with them and a range of values within which the "true" answer is likely to be [18] (as can regularly be seen with results from opinion research surveys). Not all systems present these figures explicitly in their output, but they are crucial in deciding how to use the results. Even if it is safe to assume that the training data for the system was representative of the set being presented to it in operation, this all demands that extensive testing in the intended operational environment is necessary to understand the accuracy of any predictive algorithm.

This is by no means easy to do and record comprehensively. To illustrate, in medical diagnostics, reporting on trials of predictive diagnostic systems for individual patient conditions has been found to be inadequate in the majority of cases [43]. Either the study has not been done properly and thoroughly, or the evidence has not been presented. This situation may thus occur in less experimentally rigorous fields of study and practice. Two frameworks have been created in the medical field that could be

adapted to provide a checklist for assessing how thoroughly an evaluation has been done of a prediction system for public administration [44], [45].

Finally, to reinforce the point that context and human factors are highly significant, even if an outcome is judged highly probable (or improbable) that still does not imply that a decision should follow directly from it. For example, if while a cook is foraging in the woods their app identifies a mushroom as 95% safe as opposed to 5% likely to be a deadly variety, a decision not to eat it is still the best choice. In many other situations, 95% represents a worthwhile risk [46]. The same principle applies for more complex cases [47].

Decisions in public administration can be challenged in court as discussed above. If a decision based on a statistical process is challenged, the judge may ask what was the probability that the output of the process was right. The court may conclude that in one situation the balance of probability was sufficient to support a decision; in another case it may consider that as near to certainty as possible was necessary. If a suitably-evidenced figure cannot be provided, the court's judgement is perhaps predictable.

Legality, privacy, data protection and security

Debate has primarily focused on matters of privacy and data protection in relation to AI and algorithms. However, particularly in public administration, compliance with data protection law is necessary but not sufficient: other legal or administrative principles may apply and be contravened [37], [48]. Law concerning proprietary intellectual property, patents and contract apply widely, while public sector bodies will be subject to Freedom of Information laws [27], [30].

Necessarily, in the public sector, the statutes underpinning administrative processes must be followed under the rule of law [49] — algorithms in public administration are a component of an administrative instrument that executes part of the public body's statutory functions [26]. For example it was suggested that the use of an algorithm to determine the UK 2020 exam grades was outside the lawful scope of the regulator concerned [50], [51]. The rule of law principles of predictability, consistency and equality imply that algorithms that change while in operational use may be inappropriate in public administration — inconsistent decisions could be grounds for challenge.

Freedom of Information law may allow access to the source code of the algorithm that made a contested decision, and its training data alongside other records, if they were within a public body. If they were the proprietary intellectual property of a private supplier, one would look to the field of contract law to enforce access and a public agency should consider that in procurement [27]. Buying a "black box" service is likely to be seen as a weak defence to a challenge.

Research by Cardiff University's Data Justice Lab[7] showed that several projects in English local government failed to address the basics of data protection regulations, primarily in relation to use of personal data without knowledge, consent or legal basis [7]. Inadequate safeguards for privacy and

[7] Cardiff University researchers studied three child welfare systems in England that make decisions based on previously collected data. They argue that predictive analytics tools are not neutral decision makers and question whether they should be used for any child welfare decisions, or any government processes in general. They discuss that there is a lack of transparency when it comes to the impact on the individuals affected by the systems and a conflict of views about the outsourcing of public government processes to private partnerships. One supporting case study is based on is Hackney's Children's Safeguarding Profiling System, closed in 2019.

security are also common. Given the guidance on data protection compliance and privacy impact assessments from the UK Information Commissioner's Office[8] and others, this failure is avoidable.

There is a specific case covered by the General Data Protection Regulations (GDPR) Article 22 (and Section 14 of the UK Data Protection Act 2018), where a decision is made by an algorithm without human intervention [52]. This may only be done in particular circumstances, and a subject of such a decision has a number of safeguards and remedies[9]. In public administration, a specific legal basis would almost certainly be necessary to allow this. So far, cases typically affirm that there is a human in the decision process. However, as discussed above, there may be a tendency for the automated decision to be taken by default. In such a situation where regular human agency could not be demonstrated, it is possibly that the stricter Article 22 provisions might be deemed to apply.

Security in technical and operational senses is naturally essential and must be ensured by a public body, but this paper is not the place to go into further detail on this matter.

Bias and discrimination

Algorithmic bias is a common argument when criticizing the use of predictive analytics. This is a relatively understandable and frequently justified way for critics, especially in the media, to explain to the general public why apparently dubious decisions were made [53]–[55].

Data sets represent the world in the past. Any mistakes, biases (in statistical terms this means skewed rather than prejudiced, e.g. a training dataset may predominantly contain data on white male subjects but may later be applied to wider demographics), prior environmental conditions or superseded policies will be embedded within them and the model may reproduce them to be recreated in the present. This can lead to discriminatory behaviour in an application [17], [35], [56], [57]. Whereas the point of many applications is to classify or categorise on the basis of input data, it may be socially or legally unacceptable to do so on the basis of certain explicit or implicit characteristics of the subject being classified that a model has picked up, and that is what discrimination typically means here.

A biased algorithmic process can only be shown to be discriminatory from the context of its application. Decision-making algorithms are biased by design, as by definition they create groupings and calculate outputs based on certain characteristics. An example of this could be in the application of a life insurance algorithm. An algorithm that determines how much an individual's life insurance costs could be biased against smokers, meaning individuals who smoke will have to pay more than individuals who do not. This is not discrimination if it is decided by the life insurance company that this factor should affect the outcome. However, this could be discriminatory if, for example, when looking at the outputs of the decision it is biased against people of colour. A specific social class, age or race may be more prone to being smokers and without being able to see how the algorithm comes to its decision, it is impossible to tell whether the algorithm is being biased against smokers or discriminating against a group of people for a different reason [58]. Further, it is sometimes difficult to distinguish binary measures such as smoker/non-smoker from other socio-economic factors.

[8] https://ico.org.uk/for-organisations/guide-to-data-protection/

[9] https://ico.org.uk/your-data-matters/your-rights-relating-to-decisions-being-made-about-you-without-human-involvement

An algorithm is a complicated mathematical calculation: it does not have its own morals and it does not decide to discriminate. However, the people who design, program, select data for, train and implement the algorithms can introduce their own biases or prejudices. This may well be inadvertent but is very hard to detect and eliminate. A current issue that affects this is the lack of diversity within related research areas and industry.

> *"The interests, attitudes, needs, and assumptions of the group of technology developers involved are what characterizes AI systems."* [59]

This can result in emphasising stereotypes, social segregation and limiting equal opportunities. One of the five factors used to decide whether an algorithm is fit for use, discussed earlier, is justice. Approaching justice as a value, that value being determined by the assessment of potential algorithmic bias, helps to evaluate whether the algorithm is just or not. Different sources of bias identified include training data, input design and input data (sensors, user interfaces), algorithmic requirements and goals, self-reinforcement and data correlations, and categorisations [56 p22]. Once again, forensic testing and scrutiny in the operational context is necessary to mitigate problems.

Algorithm design and training – more technical matters

Data used for training may have other problems beside bias as discussed above. Many data sets have missing values, and a statistically sound procedure for dealing with these must be applied in pre-processing the data [61]. Two or more data sets may be used alongside each other or combined, and here there is a risk of data items being defined or measured differently across them. Being aware and taking account of this is necessary.

Some data have to be transformed in order to be used in modelling, especially textual input data [62]. Text is not able to be used by models, which take numerical inputs. For example, location data such as postcodes: these need to be presented to a program in a way it can process them. They may be encoded as numbers using perhaps an index of deprivation, average rainfall, number of roundabouts, or as labels [61]. Consequently, many questions must be answered. Is geographical proximity between records important? If coded with numbers, will the computation attempt to find a relationship in the sequence? Will it try to process fractions? Will it need to work with postcodes or locations it hasn't seen before when in operation? How does it deal with infrequent data on some; does this bring in bias? Does location turn out to be a proxy for e.g. wealth? Free text such as a description of an incident requires interpretation then coding in some processable way using the improving but still imperfect techniques of natural language processing (NLP). Many potential errors and adverse implications may arise here too.

Data may be correlated, initially or following encoding, which causes problems with statistical procedures. Time series data, or anything where order matters such as speech, need special treatment. Data will have observational error attached – they will not be absolute in many cases. Statistical assumptions are often made about the probability distributions of such errors, that may not be valid. Possibly, in some settings, oddities that the system will treat as errors or outliers ("noise") are in fact indicators of some rare and important phenomenon [46].

Working out which mathematical model (e.g. a neural network) to use for a problem and particular type of data, and the statistical and mathematical methods to create an operationally useful algorithm (the result when all the modelling calculations are done) from them, is a skilled task [63]. Much like in setting

a goal, what you get is determined by what you ask for. The modeller needs to choose the type of model suited to the problem and the data – for example in machine learning there are several varieties of neural network models available (suited to different problems) with further choices of numbers, types and orders of layers of neurons. Then the training method and selection and partitioning of data have to be designed appropriately [46], [64].

Often in training a machine learning system, a measure needs to be calculated to determine when the algorithm has been found that is the "best fit" to the training data. This measure is called the "loss function", and the aim is to find its minimum value. How the loss function (and the process used to search for its minimum) are chosen will govern the algorithm that is produced. For different purposes, models, and data, different loss functions are required. This could be an obscure source of flaws that are hard to spot [65].

Mathematicians have shown that a deep learning system (a neural network with several layers of neurons) has a remarkably high chance of generating an algorithm that is unacceptable in some way (perhaps unlawful or discriminatory) if training data opens up that possibility. Such behaviour needs specific methods and tests to identify and mitigate against it [66].

"Overfitting" or "underfitting" describe unsatisfactory outcomes of training using particular data. Overfitting occurs when, while working with finite samples, the modelling discovers apparent associations that turn out not to be valid when applied to new data — in a sense, it just memorises the training data and then the algorithm can't respond to anything it doesn't recognise. Underfitting describes the situation where the model is too simple to explain a pattern in the data. This would make an algorithm a very inaccurate predictor. A model designer has to create a model that avoids such problems i.e. that is necessarily and sufficiently complex, based on the size and nature of the data set [46], [63].

It goes almost without saying that all of this needs to be converted into a computer program that is bug-free (no errors) and runs in a manageable amount of time. This is not always straightforward as machine learning programs in particular are complex things. Many of their procedures require a great deal of computing power — only recent progress in computer processor architectures and speeds by engineers, and advances in computational methods by computer scientists have made some problems solvable at all [46].

Taking the issues above as a whole, even in the event that the data collected for use in a decision-making algorithm is considered a fair set representing the real-world in the best possible way, when this data is coded and used in the algorithm, it is clearly very difficult to be sure that the result is a good match to reality and useful in public administration [67]. There are many steps between the inputs to the system and the final outcome.

> *"In the course of these complex translation processes, technological constructions of reality are generated which can stand in contrast to the constructions of the reality of individual persons or groups."* [59]

This clearly emphasises the need for a comprehensive testing process whenever an algorithm is applied to a different scenario, no matter how similar it is to the original use. In most cases, predictive analytics and machine learning algorithms are "unable to transfer their knowledge from one set of issues to

another" [59]. This may be a problem in a governmental or public administrative context as there are likely to be the same processes across different entities, e.g. councils; if there is not much local data to train the algorithm on, it could be tempting to train it with other data sets, posing a risk.

Conclusion – explainability and transparency, governance, accountability

As mentioned in the Introduction, one robust conclusion regarding the "ethical" use of algorithms is that they need to be transparent and explainable. This paper proposes that in public administration, where scrutiny and redress are inbuilt, this is already a requirement of law and the principle of the rule of law — debating "ethics" is probably superfluous.

Explainability is often used interchangeably with transparency, the latter often having the clearer added meaning of accessibility and interpretability. In almost any AI Ethics research paper, these along with governance and accountability will always come up as being very important and what is lacking the most, especially in the public sector where we argue they should be on the top of the agenda alongside accuracy [24]. Each outcome of the algorithm must be able to be explained to any stakeholder, as well as the general workings of the algorithm itself. Barriers to transparency and explainability are discussed in many research papers [17], [58], [60]. A notable one is that there is a big difference between transparency to the programmer or developer and transparency for the governed - hence the emphasis on all stakeholders having to be able to understand.

Most stakeholders will not be able to gain the insight they need into the workings of the system by its technical specification. Those stakeholders who lack the technical knowledge to understand the algorithm itself can consult experts or draw on their knowledge but often the opportunity to do this is lacking and hidden, meaning transparency and accountability is available but often diminished in practice [48]. Decision making algorithms are getting more and more complex and it is increasingly difficult to understand how some algorithms come to their decisions, even for the programmer. There is much research going into explainable AI (XAI) which is one way to try to make explainability less of an issue for situations like these [23], [68]. It aims to create programming languages, processes and reports that can explain machine learning implementations and outcomes, so they can be understood by humans. In reality, it is possible there will rarely be full explainability for the public but it is essential, should there be issues or inquiries, that there is minimum transparency that abides with the law [58], [60], [66].

> *"The most significant factor is whether the automated system uses explicit rules written by humans (generally to align with legal requirements for the relevant decision) or rules derived empirically from historic data to make inferences relevant to decisions or to predict (and thus mimic) decisions. The latter raise greater issues for transparency and accountability, particularly as newer techniques are often more complex and therefore less susceptible to human explanation."* [48]

The public sector use of algorithms, AI and machine learning presents numerous risks and challenges beyond recognized more general ones. Much hinges on the implications of the rule of law, the fundamental right of citizens to contest and seek a remedy for an administrative decision, and the role of the courts [25 s4.4]. This paper has listed many sources of risk and challenge, and thus forms a foundation upon which useful guidance may be constructed.

The spectrum of risks and issues that using an algorithm presents in this context spreads all the way from technical design to policy goals and operational use. Consequently, the understanding of the algorithm, its characteristics such as accuracy, and its use needs to be broad and deep in order to fulfil the requirements of the public accountability of officials and politicians. Public bodies cannot separate themselves from responsibility and accountability for decisions based on algorithms, no matter who else may be in their development process.

Unfortunately, every aspect mentioned inter-relates with many of the others: policy, politics, law, professional practice, human rights, data processing, statistics, mathematics, computer programming, computer science and engineering. Any public sector body considering using complex algorithms to support public administrative decisions would be well advised to conduct extensive and forensic assurance procedures within strong and accountable governance before affecting any person's real life. In the absence of full and explainable understanding of every aspect, great caution is called for.

References

[1] Z. Engin and P. Treleaven, “Algorithmic Government: Automating Public Services and Supporting Civil Servants in using Data Science Technologies,” *Comput. J.*, vol. 62, no. 3, pp. 448–460, Mar. 2019, doi: 10.1093/comjnl/bxy082.

[2] L.-M. Neudert and P. N. Howard, “Four Principles for Integrating AI & Good Governance,” 2020.

[3] British Computer Society, “The Exam Question: How Do We Make Algorithms Do The Right Thing,” 2020. [Online]. Available: https://www.bcs.org/media/6135/algorithms-report-2020.pdf.

[4] The Partnership on AI, “Report on Algorithmic Risk Assessment Tools in the U.S. Criminal Justice System,” 2019. Accessed: Sep. 23, 2020. [Online]. Available: https://www.partnershiponai.org/report-on-machine-learning-in-risk-assessment-tools-in-the-u-s-criminal-justice-system/.

[5] N. Macpherson, “Review of quality assurance of government models,” 2013. Accessed: Sep. 23, 2020. [Online]. Available: https://www.gov.uk/government/publications/review-of-quality-assurance-of-government-models.

[6] BBC News, “Home Office drops ‘racist’ algorithm from visa decisions,” 2020. https://www.bbc.co.uk/news/technology-53650758 (accessed Oct. 04, 2020).

[7] J. Redden, L. Dencik, and H. Warne, “Datafied child welfare services: unpacking politics, economics and power,” *Policy Stud.*, vol. 41, no. 5, pp. 507–526, 2020, doi: 10.1080/01442872.2020.1724928.

[8] B. Mittelstadt, “Principles Alone Cannot Guarantee Ethical AI,” *Nat. Mach. Intell.*, vol. 1, pp. 501–507, Nov. 2019, doi: 10.1038/s42256-019-0114-4.

[9] J. Morley, L. Floridi, L. Kinsey, and A. Elhalal, “From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices,” *Sci. Eng. Ethics*, no. 0123456789, 2019, doi: 10.1007/s11948-019-00165-5.

[10] A. Tsamados, L. Floridi, H. Roberts, and M. Taddeo, “The Ethics of Algorithms: Key Problems and Solutions,” 2020, Accessed: Sep. 26, 2020. [Online]. Available: https://www.academia.edu/43793187/The_Ethics_of_Algorithms_Key_Problems_and_Solutions.

[11] Department for Digital Culture Media & Sport, “Data ethics and AI guidance landscape,” *GOV.UK*, 2020. https://www.gov.uk/guidance/data-ethics-and-ai-guidance-landscape (accessed Sep. 28, 2020).

[12] Socitm, “Digital ethics: the ethical use of emerging technologies and data,” Socitm, 2020. Accessed: Sep. 26, 2020. [Online]. Available: https://socitm.net/download/digital-ethics-the-ethical-use-of-emerging-technologies-and-data/.

[13] J. Fjeld, N. Achten, H. Hilligoss, A. C. Nagy, and M. Srikumar, “Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI,” Boston, MA, 2020. Accessed: Feb. 19, 2020. [Online]. Available: http://nrs.harvard.edu/urn-3:HUL.InstRepos:42160420%0AThis.

[14] T. Hagendorff, “The Ethics of AI Ethics: An Evaluation of Guidelines,” *Minds Mach.*, vol. 30, no. 1, pp. 99–120, 2020, doi: 10.1007/s11023-020-09517-8.

[15] A. Rességuier and R. Rodrigues, “AI ethics should not remain toothless! A call to bring back the teeth of ethics,” *Big Data Soc.*, vol. 7, no. 2, 2020, doi: 10.1177/2053951720942541.

[16] M. Ryan and B. C. Stahl, “Artificial intelligence ethics guidelines for developers and users: clarifying their content and normative implications,” *J. Information, Commun. Ethics Soc.*, no. 786641, 2020, doi: 10.1108/JICES-12-2019-0138.

[17] M. Veale, M. Van Kleek, and R. Binns, “Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making,” in *Conference on Human Factors in Computing Systems - Proceedings*, Apr. 2018, vol. 2018-April, pp. 1–14, doi: 10.1145/3173574.3174014.

[18] V. Clayton, M. Sanders, E. Schoenwald, L. Surkis, and D. Gibbons, “Machine Learning in Children’s Services - Technical Report,” What Works For Children’s Social Care, London, 2020. [Online]. Available: https://whatworks-csc.org.uk/wp-content/uploads/WWCSC_technical-_report_machine_learning_in_childrens_services_does_it_work_Sep_2020.pdf.

[19] M. J. Salganik *et al.*, “Measuring the predictability of life outcomes with a scientific mass collaboration,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 117, no. 15, pp. 8398–8403, Apr. 2020, doi: 10.1073/pnas.1915006117.

[20] E. Jones and C. Safak, “Can algorithms ever make the grade?,” *Ada Lovelace Institute*, 2020. https://www.adalovelaceinstitute.org/can-algorithms-ever-make-the-grade/ (accessed Sep. 28, 2020).

[21] W. Barker, “Bridging the ethical divide: working to provide guidance,” *Socitm blog*, 2020. https://socitm.net/bridging-the-ethical-divide/ (accessed Sep. 26, 2020).

[22] L. Floridi *et al.*, “AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations,” *Minds Mach.*, vol. 28, no. 4, pp. 689–707, 2018, doi: 10.1007/s11023-018-9482-5.

[23] P. J. Phillips, C. A. Hahn, P. C. Fontana, D. A. Broniatowski, and M. A. Przybocki, “Four Principles of Explainable Artificial Intelligence,” Gaithersburg, Maryland, 2020. doi: 10.6028/NIST.IR.8312-draft.

[24] I. D. Raji *et al.*, “Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing,” in *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020, pp. 33–44, doi: 10.1145/3351095.3372873.

[25] D. Spiegelhalter, “Should We Trust Algorithms?,” *Harvard Data Sci. Rev.*, vol. 2, no. 1, pp. 1–12, Jan. 2020, doi: 10.1162/99608f92.cb91a35a.

[26] P. Waller and V. Weerakkody, “Digital Government: overcoming the systemic failure of transformation,” Brunel University London, 2016. [Online]. Available: http://bura.brunel.ac.uk/handle/2438/12732.

[27] G. N. La Diega, “Against the Dehumanisation of Decision-Making – Algorithmic Decisions at the Crossroads of Intellectual Property, Data Protection, and Freedom of Information,” *JIPITEC*, vol. 9, no. 1, May 2018, Accessed: Sep. 17, 2020. [Online]. Available: https://www.jipitec.eu/issues/jipitec-9-1-2018/4677.

[28] V. Eubanks, *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New

York: Picador, 2019.

[29] D. Hogan-Doran, “Computer says ‘ no ’: automation, algorithms and artificial intelligence,” in *UNSW Public Sector Law & Governance Seminar*, 2017, p. 39, [Online]. Available: https://www.academia.edu/38540521/Computer_says_no_automation_algorithms_and_artificial_intelligence_in_Government_decision-making?email_work_card=view-paper.

[30] M. Oswald, “Algorithm-assisted decision-making in the public sector: framing the issues using administrative law rules governing discretionary power,” *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.*, vol. 376, Sep. 2018, doi: 10.1098/rsta.2017.0359.

[31] M. Tisné and M. Schaake, “The Data Delusion: Protecting Individual Data is Not Enough When the Harm is Collective.” Luminate, pp. 1–12, 2020, [Online]. Available: https://cyber.fsi.stanford.edu/publication/data-delusion.

[32] A. Elbanna and J. Engesmo, “A-level results: why algorithms get things so wrong – and what we can do to fix them,” *The Conversation*, 2020. https://theconversation.com/a-level-results-why-algorithms-get-things-so-wrong-and-what-we-can-do-to-fix-them-142879 (accessed Oct. 04, 2020).

[33] M. T. Cicero, *De officiis*. Rome, 44BC.

[34] C. E. Church and A. J. Fairchild, “In Search of a Silver Bullet: Child Welfare’s Embrace of Predictive Analytics,” *Juv. Fam. Court J.*, vol. 68, no. 1, pp. 67–81, Mar. 2017, doi: 10.1111/jfcj.12086.

[35] P. Gillingham, “Decision Support Systems, Social Justice and Algorithmic Accountability in Social Work: A New Challenge,” *Practice*, vol. 31, no. 4, pp. 277–290, Aug. 2019, doi: 10.1080/09503153.2019.1575954.

[36] Parliamentary Office of Science and Technology, “Interpretable Machine Learning,” UK Parliament, London, 2020. [Online]. Available: https://post.parliament.uk/research-briefings/post-pn-0633/.

[37] E. Keddell, “The ethics of predictive risk modelling in the Aotearoa/New Zealand child welfare context: Child abuse prevention or neo-liberal tool?,” *Crit. Soc. Policy*, vol. 35, no. 1, pp. 69–88, Feb. 2015, doi: 10.1177/0261018314543224.

[38] A. Pithouse, K. Broadhurst, C. Hall, S. Peckover, D. Wastell, and S. White, “Trust, risk and the (mis)management of contingency and discretion through new information technologies in children’s services,” *J. Soc. Work*, vol. 12, no. 2, pp. 158–178, Mar. 2012, doi: 10.1177/1468017310382151.

[39] A. Campolo and K. Crawford, “Enchanted Determinism: Power without Responsibility in Artificial Intelligence,” *Engag. Sci. Technol. Soc.*, vol. 6, no. 0, p. 1, Jan. 2020, doi: 10.17351/ests2020.277.

[40] E. Regnier, “How to interpret coronavirus news in light of false negatives,” *Analytics*, no. April, pp. 1–11, 2020, doi: 10.1287/lytx.2020.03.01.

[41] A. Engler, “A guide to healthy skepticism of artificial intelligence and coronavirus,” 2020. Accessed: Sep. 23, 2020. [Online]. Available: https://www.brookings.edu/research/a-guide-to-healthy-skepticism-of-artificial-intelligence-and-coronavirus/.

[42] V. Sena, A. Crispe, S. Smith, O. Sergushova, and L. A. Toderas, “Application of predictive analytics

in social care: tools to support decision making," in *Catalyst Conference*, 2019, no. June, [Online]. Available: https://www.essex.ac.uk/-/media/documents/research/catalyst/catalyst-workshope-analytics-social-care.pdf.

[43] M. Nagendran *et al.*, "Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies," *BMJ*, p. 368:m689, Mar. 2020, doi: 10.1136/bmj.m689.

[44] G. S. Collins, J. B. Reitsma, D. G. Altman, and K. G. M. Moons, "Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement," *BMJ*, vol. 350, Jan. 2015, doi: 10.1136/bmj.g7594.

[45] "Probast." http://www.probast.org/scope/ (accessed Sep. 23, 2020).

[46] A. Zhang, Z. C. Lipton, M. Li, and A. J. Smola, *Dive into Deep Learning*. 2020.

[47] C. Williams and J. Khim, "Utility Functions," *Brilliant Math & Science Wiki*. https://brilliant.org/wiki/utility-functions/ (accessed Sep. 23, 2020).

[48] M. Zalnieriute, L. Burton, J. Boughey, L. Bennett Moses, and S. Logan, "From Rule of Law to Statute Drafting: Legal Issues for Algorithms in Government Decision-Making," *SSRN Electron. J.*, 2019, doi: 10.2139/ssrn.3380072.

[49] M. Zalnieriute, L. B. Moses, and G. Williams, "The Rule of Law and Automation of Government Decision-Making," *Mod. Law Rev.*, vol. 82, no. 3, 2019, [Online]. Available: https://www.academia.edu/38515429/The_Rule_of_Law_and_Automation_of_Government_Decision-Making.

[50] J. Elgot and R. Adams, "Ofqual exam results algorithm was unlawful, says Labour," *The Guardian*, 2020. https://www.theguardian.com/education/2020/aug/19/ofqual-exam-results-algorithm-was-unlawful-says-labour (accessed Sep. 28, 2020).

[51] C. Kind, J. Tennison, R. Coldicutt, S. L. Harris, and C. Crider, "Five data-ethics and AI experts on... what we can learn from the qualifications algorithm fiasco," *Civil Service World*, 2020. https://www.civilserviceworld.com/professions/article/data-ethics-ai-experts-on-what-we-can-learn-from-the-qualifications-algorithm-fiasco (accessed Sep. 28, 2020).

[52] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR," *Harv. J. Law Technol.*, vol. 31, no. 2, 2018, [Online]. Available: https://jolt.law.harvard.edu/assets/articlePDFs/v31/Counterfactual-Explanations-without-Opening-the-Black-Box-Sandra-Wachter-et-al.pdf.

[53] M. A. Gianfrancesco, S. Tamang, J. Yazdany, G. Schmajuk, S. Shiboski, and L. Mackey, "Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data," *JAMA Intern Med*, vol. 178, no. 11, pp. 1544–1547, 2018, doi: 10.1001/jamainternmed.2018.3763.

[54] K. S. Gill, "Prediction paradigm: the human price of instrumentalism," *Ai Soc.*, no. 0123456789, p. 93, 2020, doi: 10.1007/s00146-020-01035-6.

[55] K. Miller, "A Matter of Perspective: Discrimination, bias and inequality in AI," in *Closer To The Machine - Technical, social, and legal aspects of AI*, C. Bertram, A. Gibson, and A. Nugent, Eds. Melbourne: Office of the Victorian Information Commissioner, 2019, pp. 23–39.

[56] V. U. Prabhu and A. Birhane, “Large image datasets: A pyrrhic win for computer vision?,” 2020, [Online]. Available: http://arxiv.org/abs/2006.16923.

[57] A. Torralba, R. Fergus, and B. Freeman, “80 Million Tiny Images,” 2020. https://groups.csail.mit.edu/vision/TinyImages/ (accessed Sep. 23, 2020).

[58] T. van Nuenen, X. Ferrer, J. M. Such, and M. Cot, “Transparency for whom? Assessing discriminatory AI,” *Computer (Long. Beach. Calif).*, pp. 1–9, 2020, [Online]. Available: https://kclpure.kcl.ac.uk/portal/en/publications/transparency-for-whom-assessing-discriminatory-ai(3f16c127-4ce7-44ba-9478-19dbaf333a51).html.

[59] T. Hagendorff and K. Wezel, “15 challenges for AI: or what AI (currently) can’t do,” *AI Soc.*, vol. 35, no. 2, pp. 355–365, 2020, doi: 10.1007/s00146-019-00886-y.

[60] D. Drinka, K. Voge, and M. Y.-M. Yen, “From Principles to Practice,” pp. 177–195, 2011, doi: 10.4018/978-1-59140-673-0.ch011.

[61] J. Brownlee, “Tour of Data Preparation Techniques for Machine Learning,” *Machine Learning Mastery*, 2020. https://machinelearningmastery.com/data-preparation-techniques-for-machine-learning/ (accessed Sep. 23, 2020).

[62] S. A. N. Alexandropoulos, S. B. Kotsiantis, and M. N. Vrahatis, “Data preprocessing in predictive data mining,” *Knowl. Eng. Rev.*, vol. 34, 2019, doi: 10.1017/S026988891800036X.

[63] S. Sengupta *et al.*, “A review of deep learning with special emphasis on architectures, applications and recent trends,” *Knowledge-Based Syst.*, vol. 194, p. 105596, Apr. 2020, doi: 10.1016/j.knosys.2020.105596.

[64] J. M. Hofman, A. Sharma, and D. J. Watts, “Prediction and explanation in social systems,” *Science (80-.).*, vol. 355, no. 6324, pp. 486–488, Feb. 2017, doi: 10.1126/science.aal3856.

[65] J. Brownlee, “Loss and Loss Functions for Training Deep Learning Neural Networks,” *Machine Learning Masteryg Mastery*, 2019. https://machinelearningmastery.com/loss-and-loss-functions-for-training-deep-learning-neural-networks/ (accessed Sep. 23, 2020).

[66] N. Beale, H. Battey, A. C. Davison, and R. S. MacKay, “An unethical optimization principle,” *R. Soc. Open Sci.*, vol. 7, no. 7, p. 200462, 2020, doi: 10.1098/rsos.200462.

[67] D. K. Citron, “Technological Due Process,” *Washingt. Univ. Law Rev.*, vol. 85, no. 6, pp. 1249–1313, 2008, doi: 10.2139/ssrn.1012360.

[68] T. Spinner, U. Schlegel, H. Schäfer, and M. El-Assady, “ExplAIner: A Visual Analytics Framework for Interactive and Explainable Machine Learning,” *IEEE Trans. Vis. Comput. Graph.*, vol. 26, no. 1, pp. 1064–1074, Jan. 2020, doi: 10.1109/TVCG.2019.2934629.

[69] European Commission, “Communication - 2020 Rule of law report - the rule of law situation in the European Union,” Brussels, 2020. Accessed: Oct. 05, 2020. [Online]. Available: https://ec.europa.eu/info/files/2020-rule-law-report-rule-law-law-situation-european-union_en.

[70] BBC News, “A-levels and GCSEs: How did the exam algorithm work?,” 2020. https://www.bbc.co.uk/news/explainers-53807730 (accessed Sep. 23, 2020).