Connecting diverse public sector values with the procurement of machine learning systems.

Michael Veale

Department of Science, Technology, Engineering and Public Policy (STEaPP)
University College London
Gower Street, London WC1E 6BT, UK
m.veale@ucl.ac.uk

Abstract

Increasing interest in using machine learning systems for decision making and support in the public sector has raised questions as to how these technologies can be designed, implemented and managed responsibly. This short discussion paper describes some relevant social and technical potentials and perils of machine learning by relating them to different groups of public sector values outlined in the public administration literature. Practitioners may find this structure useful to help them understand different dimensions of responsibility they may wish to consider if they are considering using these technologies, and how they link to developing work and tools in the field.

Keywords— machine learning; public policy; public administration; public procurement.

1 Introduction

Machine learning (ML) has emerged as a promising but controversial technology in recent years. At its core is the ability to algorithmically model useful patterns in large datasets that are difficult for humans to spot. These models can then be used for tasks such as predicting or structuring data. Following successful applications in the business world and several high-profile public demonstrations, there has been considerable interest in applying it to public sector challenges.

Yet precisely how to do this is far from clear. These technologies are usually procured or developed either in a consortium or by a contractor or vendor. Guidance and 'best practices' for responsible procurement in this space are thin on the ground. This short discussion paper aims to provide helpful structure to this issue through discussing three groups of public sector values identified in the public administration literature, and using them to group discussion of potential, perils and emerging pathways around these technologies.

In a classic paper, Hood (1991), drawing on previous work in management values and normative aims in the public sector, distinguishes between three sets of core values in public management.

Sigma-type values are characterised by 'keeping it lean and purposeful'. Central to the trend towards so-called *New Public Management*, these concern matching resources to well-defined tasks in a 'competent and sparing fashion'. A *sigma*-worldview sees success in terms of efficiency, and failure in terms of waste and confusion.

Theta-type values are characterised by the prioritisation fairness and honesty. A *theta*-driven success is one where a legitimate process has achieved a proper discharge of public duty, avoiding abuses of office, distortions, inequities, biases and regulatory capture. Naturally, this is contentious and difficult to pin down, but it is this precise broad and changing nature of the value-set that characterises it.

Lambda-type values consider broader systemic risk, resilience, robustness and adaptation of public sector systems. For those with a *lambda*-affinity, failure come with collapse or breakdown, and their day-to-day currency is systemic confidence.

These three flavours of value can be used as a lens to consider the hopes piled onto machine learning systems, and the particular types of failure they could engender. Below, they are treated in order. The topics discussed within are by no means complete, due to both space limitations and the breadth of the field, but they aim to sample core issues in this space and the type of approaches suggested to deal with them.

2 Sigma σ: lean, purposeful ML

Potential of ML Given the extensive public holdings of and privileged access to a range of data sources, the potential to extract new value from administrative data has generated much excitement. This value might include richer, cheaper or more regular statistics at new temporal or spatial resolutions (Bańbura, Giannone, Modugno, & Reichlin, 2012; Smith, Quercia, & Capra, 2013; Smith-Clarke, Mashhadi, & Capra, 2014; Struijs, Braaksma, & Daas, 2014), and a range of new predictive capabilities in fields such as security and infrastructure (Bean, 2015). Assuming that tasks can be delineated and well-defined, decision-making and support systems might increase the effectiveness and efficiency of the work of analysts, or might enable staff reduction or redeployment through automating rote tasks, such as mail redirection (Wilcocks & Lacity, 2016).

Perils of ML Failure in the sigma-worldview is primarily seen in terms of wastage and inefficiency. The information systems literature often discusses this type of failure — not clearly defined, but something you'd supposedly know when you see it. The sources of kind of failures in traditional IT projects are thought to include issues of staffing, organisational politics, poor requirements considerations, unrealistic planning, badly identified needs, and broader issues of complexity and ability (Wallace, Keil, & Rai, 2004). Machine learning projects are never far away from these perils, partly due to their contentious and cross-departmental natures, their novelty, and both local and sector-wide issues of resource and expertise scarcity.

Particular liabilities specific to ML also endanger its 'leanness'. While all IT requires maintenance, we have few operational examples of how this maintenance might be undertaken in consequential predictive systems. 'In the wild', datasets are usually not drawn from a static population but a changing one. Some of the patterns and correlations involving areas predictive systems have currently been developed for, such as tax fraud, child abuse or food safety, are moving targets the links between variables being used to predict and variables trying to be predicted is not static. Several different types of movement have been distinguished by researchers, and are broadly called 'concept drift' or 'dataset shift' (Moreno-Torres, Raeder, Alaiz-Rodríguez, Chawla, & Herrera, 2012), and a range of methods have been developed to cope with them. As noted by Gama, Žliobaitė, Bifet, Pechenizkiy, and Bouchachia (2013) in their review of the field, a single technological solution to this issue is not forthcoming, and a priority for future robust systems will be better knowledge of how to integrate expertise as checks and balances into these models. Expertise is expensive and scarce, and models that can integrate this into technical analyses whilst remaining realistic in their requirements are sorely needed.

One step further than this is a phenomenon especially crucial to the public sector. In a laboratory setting, or within a 'price-taking' company, there is a core assumption that incoming data is not a function of your model or its predictions. This assumption becomes questionable in the public sector, which is the sole or near-monopoly supplier of many services with significant resource and privilege. A system used to help deploy police officers is likely to affect the distribution and nature of crime in some way that is the entire point. It might also affect a whole array of other features: public trust, perceptions of safety or the housing market. Concept drift is already difficult and likely to be expensive to manage. Emergent complexity through these kind of feedback processes is likely to compound this concern, and remarkably little work has been undertaken in this field.

Legal-political liabilities around ML systems might also increase their baggage. While rarely operationally tested, variously worded rights to be given the 'logic of processing' of a (significant, fully-automated) decision have existed since the EU's Data Protection Directive 95/46/EC, and have been reiterated in the recent General Data Protection Regulation 2016/679. While the high bar needed to trigger this clause is rarely met, particularly not for public sector decisions, it is not unimaginable that future governance mechanisms include Freedom of Information-style burdens or required alternative human-processing routes on organisations using machine learning systems, incurring significant additional expense. Contentious public technologies bring specific investment risks.

3 Theta θ: honest, fair ML

Potential of ML While these technologies are often portrayed in the media as the antithesis towards honesty and fairness, *theta*-types do have reason for positivity. Firstly, the use of ML systems promotes a particular type of explicit problem definition. Choosing both the (usually labelled) data used to train the systems and the performance metrics and loss functions used to evaluate them may be an opportunity to make values behind the decision system more explicit although defining a problem narrowly also brings its own troubles. In areas where there is concern that decentralised human discretion is causing fairness issues in decision-making, it might be that an ML system, while far

from bias free, is more feasible to use to probe for these.

The assessment of the performance of an ML system also demands, both operationally and politically, significant collection of data for evaluation purposes for example, to prevent problems with concept drift. Investments in analytics capacity and in data collection for monitoring and evaluation generally underfunded aspects of policy programmes might have the spillover effect of being useful for undertaking audits of fairness issues difficult or impossible without this investment. Programme evaluation has generally been thought as a crucial operational component of public trust and legitimacy (Vedung, 1997).

Perils of ML Discrimination, fairness and bias have occupied much of the recent interest in algorithms from social scientists, as well as that of a significantly smaller proportion of technically-oriented scholars.

Possible sources of ML fairness and discrimination issues are broad, even when somewhat divorcing the technology from its use in practice. They can relate to the data, which could be imbalanced, aggregated or disaggregated at scales that exacerbate fairness issues, could omit important contextual variables, or be cleaned or classified problematically. Processing can also be important: different analytic methods are more or less adept at picking up particular types of patterns and make different structural assumptions about relations in the world. A *random forest* algorithm can pick up certain types of synergistic effects between variables, while linear regression cannot do so automatically. They can also relate broadly to the entire predictive approach: is it fair to judge future decisions based on past patterns? At what point in time should an algorithm 'forget' the more distant history of an area or a person?

Techniques exist both to assess the discrimination in predictive systems and to 'scrub' it out. Some attempt to process or 'massage' biased data to avoid or limit discrimination, while others modify algorithms to be 'fairness-aware'. To mathematically operationalise concepts such as fairness and discrimination, formal definitions, primarily rooted in law, are taken (Hajian, Domingo-Ferrer, Monreale, Pedreschi, & Giannotti, 2015). Nevertheless, what is seen as a legitimate action in society is often very different from what is legally permissible, so it seems wise to move towards considering how technical approaches might support more abstract notions of fairness.

Many challenges remain. Discrimination may not manifest simply, but through combinations of variables. Even when women are treated the same as men, older women, or perhaps older women in a certain area, may be treated differently from their male counterparts. In a dataset with many variables, particularly continuous ones, it becomes difficult to locate synergistic discrimination. Omitting a particular sensitive variable is no guarantee of discrimination avoidance combinations of other, less sensitive variables might 'proxy' the effect of the sensitive one, if it was predictively useful.

Methods to remove this type of discrimination tend to require the sensitive variables be collected, stored and processed. This is at tension with calls for *data minimisation* in legal frameworks such as that of the EU, and potentially with public trust more broadly. Only using this for anti-discriminatory purposes provides no guarantee about how these systems will be publicly perceived, and these perceptions might damage the *theta* obligation of public duty, trust and legitimacy. Issues of trust may also enter into the accuracy–fairness trade-offs that are necessary to navigate in discrimination-aware data science.

The importance of perception to the success and acceptance of decision support systems has been long discussed in the expert systems literature (Teach & Shortliffe, 1984). The latest calls for intelligibility of algorithms (e.g. Pasquale, 2015) and more nuanced treatments of their opacity (Burrell, 2016) add to a considerable history of thought about how best to explain and allow the interpretation of decision-support systems, much of which has received inadequate treatment in recent years.

Work on interpretability of expert systems sought to make simplified *traces* of the reasoning process, or explanations from different angles of perspectives. This might include *how* analysis was carried out, or *why* a certain strategy was used. As thought developed explanation was increasingly seen as an independent problem that could only be tacked by decoupling it from the logic of reasoning (Wick & Thompson, 1992). This would entail making a model for explanation, a model for prediction necessary as the more predictively powerful machine learning systems available today cannot represent knowledge structures in an explicit, declarative way. While there has been some work trying to get both explanation and powerful prediction from the same models, more seems forthcoming from somewhat or wholly *pedagogical* or *model-agnostic* approaches, which treat a black-boxed model as something that transforms inputs into outputs, attempting to model its main logics without 'opening it up' (Andrews, Diederich, & Tickle, 1995; Tickle, Andrews, Golea, & Diederich, 1998). Yet explaining logic semantically is often problematic in a 'big data' world transparency falls away as this explanation often becomes too unwieldy

for a human to grasp (Nissenbaum, 2011). The number of *if-then* clauses needed to 'explain' a neural network usefully is often too large to be helpful. Consequently, more recent work has tried to go beyond expressing these logics in text, using visualisation and media to help 'explain' local logics of predictions from high dimensional data, such as images (Ribeiro, Singh, & Guestrin, 2016). Yet there is a dearth of studies understanding how these systems help interpretability in practice, and what effects this has, both on *theta* values but also on areas of interest to *sigma* and *lambda* types.

4 Lambda λ: robust and resilient ML systems

Potential of ML *Lambda*-values are potentially the most underdeveloped in practice, particularly in terms of analytical and institutional frameworks for integrating them into policymaking. Complex adaptive systems theory has pointed to the fragility of static sociotechnical systems, so introducing adaptive elements such as machine learning may have potential to create more fluid responses to rapid change. The ability to turn to automated systems may also allow better coping in extreme levels of demand, or adverse situations. Yet *lambda*-style values imply a wariness of silver bullets in the context of a broader, interlinked systems perspective.

Perils of ML Machine learning systems are complex, adaptive, and interplay with many neighbouring sociotechnical systems. The array of potential impacts and feedback effects presents a steep challenge for *lambda*-type values.

As most proposed ML systems in the public sector are decision-support systems, there is a large and flexible gap between prediction and action. This gap can be both a check and balance on the quality of the model, for example with the *sigma* approach to accuracy and the *theta* approaches to promoting fairness, but it can also introduce new issues. A considerable body of research points to the issues both laypeople and professionals have in interpreting probability (e.g. Gigerenzer, Gaissmaier, Kurz-Milcke, Schwartz, & Woloshin, 2007) and the inconsistency of understanding of semantic representations of uncertainty (Wallsten, Fillenbaum, & Cox, 1986). *Automation bias* refers to times where decision-makers either fail to question or take the advice of automated systems (Skitka, Mosier, & Burdick, 1999), and a range of research highlights areas where we are too forgiving of errors (Dijkstra, 1999) or judge systems too harshly when they err (Dietvorst, Simmons, & Massey, 2015; Dzindolet, Peterson, Pomranky, Pierce, & Beck, 2003). This work also highlights the benefits of making users aware of situations where machine learning systems might be more error prone (Dzindolet et al., 2003), which links to broader lessons in *lambda*-type systemic areas about the crucial roles of informational feedback in system success.

A further *lambda* concern surrounds the sociotechnical environment of a machine learning system. ML systems are highly reliant on a range of organisational activities, such as data collection, data clean, model and interface building and evaluation, and so on. Especially where the individuals performing functions such as data collection are also the ones using the decision-support system using the data — such as police officers on patrol — possibilities for feedback and gaming emerge. In the police sector, gaming already includes not recording crime ('cuffing'), attributing crimes to willing offenders without increasing sentences ('nodding'), evidence fabrication ('stitching', usually rare), and concentrating on easy-to-solve crimes for statistical purposes ('skewing') (Patrick, 2009). These new dynamics might have unforeseen consequences on a model's evolution and functioning.

Gaming can also occur in data cleaning. In Mid Staffordshire NHS Foundation Trust, a data coder noticed that some patients who enter hospital with a fractured hip but later die of pneumonia caught inside can have their primary diagnosis recoded to the latter. By recoding this way to play to institutional incentives, it seemed a patient with a hip fracture was five times *less* likely to die if admitted to Mid Staffordshire than on average, despite scandalous hospital conditions (Hammond, 2013; Hawkes, 2013). While at the time of writing there are no public ML-driven scandals of this magnitude, the importance of these neighbouring systems should not be underestimated.

ML systems may have further impacts on the internal workings of a public sector process. Are decision-making systems resilient to the failure or speedy retirement of ML decision-support software? What effect does the use of this type of software have on the creation, retention and dissemination of the tacit knowledge which allows analysts to do their work without them? From a social lens, there may be aspects of these technologies which reduce certain capacities as they increase others.

Issues concerning *lambda*-types originate from without as well as within. Machine learning systems have cybersecurity vulnerabilities not experienced by other types of software. Some attacks can attempt to change the learning model itself, potentially trying to compromise its integrity or make it so inaccurate as to be useless. Other attacks are exploitative of

misclassifications and the probabilistic nature of the technology, potentially trying to manipulate specific decisions or to gain access to private information (Barreno, Nelson, Joseph, & Tygar, 2010; Huang, Joseph, Nelson, Rubinstein, & Tygar, 2011). While the field of *adversarial machine learning* is relatively young, it becomes considerably more important the larger the uptake of these systems are in and near critical areas of government.

5 Concluding remarks

The discussed value-sets are far from the only relevant perspectives on the benefits and dangers of ML technologies. However, as three common public sector viewpoints they represent a basic point of departure for discussions of some of the issues needing operational consideration. There is still considerable work to be done. Below, I briefly highlight two directions of research and practice I see as key: *practices of responsibility* and *responsibility in practice*.

Practices of responsibility Few empirical studies have examined how machine learning is designed, deployed and managed in real, high-stakes environments. Important lessons might be learned from other cases abour what has been devised and what has seen success or failure. Which of the above challenges (or others) were considered, and which were acted upon? How did cost and time constraints, as well as existing frameworks such as Privacy Impact Assessments, shape the procurement process? What institutional structures were used or developed to explore these technologies, and what were the roles of stakeholders and different types of expertise within them? Researchers need to go beyond anecdotal evidence or evidence from the media. Practitioners here should seek out past cases in their field and attempt to form or join networks to share knowledges and generate practices that can be tested and validated. Both should try to make time to record and reflect on projects they are involved in, and consider how they relate to other public sector aims or academic disciplines.

Responsibility in practice In the sections above, a range of the technical approaches that have been proposed to ameliorate issues in this space are touched upon. While some of these have been developed into open source packages, accounts of them being deployed in practice are sparse. There is a strong need for reflective studies of how these technologies are used and might be used, both to promote them in more wary and less experimental environments and to understand where further research and development might be needed. Researchers need to go beyond theoretical validation to 'in-the-wild' evaluation, and practitioners should seek out new sociotechnical tools to further evaluate dimensions of responsibility they may have initially overlooked.

It is easy to overstate issues around machine learning in the public sector, or to make them seem more novel than they really are. In many ways, new ICT technologies do not introduce wholly novel ethical concerns the public sector has always had the ability to commission expensive technological failures, to discriminate unfairly and to lose public trust, and to fail systemically. Yet on an operational level, ML *does* raise new challenges, and there is a real need for better engagement between scientists and public administrators to develop replicable practices and leading examples of responsible processes in this emerging and challenging space.

Acknowledgements

This work has been supported by the UK's Engineering and Physical Sciences Research Council (EPSRC) through a Doctoral Training Partnership grant (EP/M507970/1).

References

Andrews, R., Diederich, J., & Tickle, A. B. (1995). Survey and critique of techniques for extracting rules from trained artificial neural networks. *Knowledge-Based Systems*, *8*(6), 373–389.

Bańbura, M., Giannone, D., Modugno, M., & Reichlin, L. (2012). Now-casting and the real-time data flow. *European Central Bank Working Paper Series*(1564). Retrieved from `https://www.ecb.europa.eu/pub/pdf/scpwps/ecbwp1564.pdf`

Barreno, M., Nelson, B., Joseph, A. D., & Tygar, J. D. (2010). The security of machine learning. *Machine Learning*, *81*(2), 121–148. doi: 10.1007/s10994-010-5188-5

Bean, C. (2015). *Independent review of UK economic statistics: Interim report*. London: HM Government. Retrieved from `https://www.gov.uk/government/publications/independent-review-of-uk-economic-statistics-interim-report`

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, *3*(1). doi: 10.1177/2053951715622512

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algo-

rithms after seeing them err. *Journal of Experimental Psychology: General*, *144*(1), 114. doi: 10.1037/xge0000033

Dijkstra, J. J. (1999). User agreement with incorrect expert system advice. *Behaviour & Information Technology*, *18*(6), 399–411. doi: 10.1080/014492999118832

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International Journal of Human-Computer Studies*, *58*(6), 697–718. doi: 10.1016/S1071-5819(03)00038-7

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

Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L. M., & Woloshin, S. (2007). Helping doctors and patients make sense of health statistics. *Psychological Science in the Public Interest*, *8*(2), 53–96.

Hajian, S., Domingo-Ferrer, J., Monreale, A., Pedreschi, D., & Giannotti, F. (2015). Discrimination- and privacy-aware patterns. *Data Mining and Knowledge Discovery*, *29*(6). doi: 10.1007/s10618-014-0393-7

Hammond, P. (2013). Return to the killing fields: A chronicle of deaths foretold. *Private Eye*, *1334*, 11–13.

Hawkes, N. (2013). How the message from mortality figures was missed at Mid Staffs. *BMJ*, *346*, f562. doi: 10.1136/bmj.f562

Hood, C. (1991). A public management for all seasons? *Public Administration*, *69*, 3–19. doi: 0.1111/j.1467-9299.1991.tb00779.x

Huang, L., Joseph, A. D., Nelson, B., Rubinstein, B. I. P., & Tygar, J. D. (2011). Adversarial machine learning. doi: 10.1145/2046684.2046692

Moreno-Torres, J. G., Raeder, T., Alaiz-Rodríguez, R., Chawla, N. V., & Herrera, F. (2012). A unifying view on dataset shift in classification. *Pattern Recognition*, *45*(1), 521–530. doi: 10.1016/j.patcog.2011.06.019

Nissenbaum, H. (2011). A contextual approach to privacy online. *Dædelus*, *140*(4), 32–48. doi: 10.1162/DAED_a_00113

Pasquale, F. (2015). *The black box society: The secret algorithms that control money and information*. Cambridge, MA: Harvard University Press.

Patrick, R. (2009). *Performance Management, Gaming and Police Practice: A study of changing police behaviour in England and Wales during the era of New Public Management* (Doctoral dissertation, University of Birmingham). Retrieved from `http://etheses.bham.ac.uk/534/1/Patrick09PhD.pdf`

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *"Why Should I Trust You?": Explaining the Predictions of Any Classifier*. Retrieved from arXiv (1602.04938).

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

Smith, C., Quercia, D., & Capra, L. (2013). Finger on the pulse: Identifying deprivation using transit flow analysis. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work* (pp. 683–692).

Smith-Clarke, C., Mashhadi, A., & Capra, L. (2014). Poverty on the cheap: Estimating poverty maps using aggregated mobile communication networks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 511–520).

Struijs, P., Braaksma, B., & Daas, P. J. (2014). Official statistics and Big Data. *Big Data & Society*, *1*(1), 1–6. doi: 10.1177/2053951714538417

Teach, R. L., & Shortliffe, E. H. (1984). Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project. In B. G. Buchanan & E. H. Shortliffe (Eds.), *Rule-based expert systems* (pp. 635–652). Reading, MA: Addison Wesley.

Tickle, A. B., Andrews, R., Golea, M., & Diederich, J. (1998). The truth will come to light: directions and challenges in extracting the knowledge embedded within trained artificial neural networks. *IEEE Transactions on Neural Networks*, *9*(6), 1057–1068. doi: 10.1109/72.728352

Vedung, E. (1997). *Public policy and program evaluation*. New Brunswick, NJ: Transaction Publishers.

Wallace, L., Keil, M., & Rai, A. (2004). Understanding software project risk: a cluster analysis. *Information & Management*, *42*(1), 115–125. doi: 10.1016/j.im.2003.12.007

Wallsten, T. S., Fillenbaum, S., & Cox, J. A. (1986). Base rate effects on the interpretations of probability and frequency expressions. *Journal of Memory and Language*, *25*(5), 571–587. doi: 10.1016/0749-596X(86)90012-4

Wick, M. R., & Thompson, W. B. (1992). Reconstructive expert system explanation. *Artificial Intelligence*, *54*(1-2), 33–70. doi: 10.1016/0004-3702(92)90087-E

Wilcocks, L. P., & Lacity, M. C. (2016). *Service automation*. Stratford-upon-Avon: Steve Brookes Publishing.