Algorithmic Regulation

David Demortain, Bilel Benbouzid

carr
centre for analysis
of risk and regulation

KING'S College LONDON

Algorithmic Regulation

Leighton Andrews, Bilel Benbouzid, Jeremy Brice, Lee A. Bygrave, David Demortain, Alex Griffiths, Martin Lodge, Andrea Mennicken, Karen Yeung

LSE THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

DISCUSSION PAPER No: **85**
DATE: **September 2017**

Algorithmic regulation

Contents

Published by the Centre for Analysis of Risk and Regulation at the
London School of Economics and Political Science
Houghton Street
London WC2A 2AE
UK

2049-2718

Algorithmic regulation

Algorithmic regulation has become a central theme in contemporary policy discussions (see generally Yeung 2017). The broader social implications of our increasing reliance on algorithms in daily life have attracted considerable interest in recent years, especially with the rising awareness of the power of 'big data' and predictive analytics. One of the most vivid examples is the widespread concern about the use of algorithms to manipulate information and therefore affect political life, especially at election time, at least since the US elections. In the economy, the role of cryto-currencies is seen as an important development, in which cryptographic algorithms play a critical role, as is the use of algorithms to facilitate the supply and demand of services across the so-called 'gig economy'.

At the same time, governments have been keen to harness the power of algorithms to inform decision making in a range of policy spheres, including the use of algorithms to optimise resource allocation decisions, and to do so pre-emptively, exemplified in the increasing popularity of so-called 'predictive policing'. Concerns about the need and importance to hold algorithmic power to account invariably extends to the role of law generally, and data protection law in particular, to secure algorithmic accountability, yet it is far from clear that existing mechanisms are up to the task.

In view of the pervasive influence of the algorithm in economic, political and social life, the world of regulation scholarship has also become increasingly interested in the implications of algorithms for and in regulation. The papers in this collection were initially presented during a joint workshop between King's College London's Centre for Law Technology, Ethics & Society (TELOS) and the London School of Economics' Centre for Analysis of Risk and Regulation (**carr)** held in early July 2017. Professor Helen Nissenbaum from New York University was our distinguished international guest. The workshop was organised to bring together a range of international scholars from different disciplinary backgrounds to discuss emerging themes in algorithmic regulation. This collection brings together a revised selection of papers that provided the basis for discussion during the workshop.

The workshop funding was provided by **carr**'s ESRC funded 'Regulation in Crisis?' seminar series and by The Dickson Poon School of Law, in supporting the work of the Centre for Technology, Ethics, Law & Society (TELOS) at King's College London.

Martin Lodge and Karen Yeung

Reference

Yeung, K. (2017) 'Algorithmic regulation: a critical interrogation', *Regulation & Governance*, doi: 10.1111/rego.12158.

The importance of regulation *of* and *by* algorithm

Martin Lodge and Andrea Mennicken

The regulation of and by algorithms has become of growing interest to students of regulation, coinciding with the related interest in open and big data. Early contributions on the potential implications of the rise of algorithmic regulation focused on the interaction between social and technical determinants. These discussions considered whether the rise of algorithmic regulation and new information technologies represented a fundamental (mostly benevolent) change in opportunities for citizens and states (and opportunity structures), whilst others pointed to the likely reinforcement of existing power structures (such as the detecting powers of states), or the rise of new unregulated and private sources of surveillance. Yet others noted the likely complexification effects of the use of computerised algorithms in generating new types of unintended consequences.

It is therefore unsurprising that algorithmic regulation has become a growth industry in the study of changing modes of surveillance (often under the 'risk-based' label) by state actors as well as in the disquisitions on the rise of new forms of corporate power, based on the economic value of 'data'. Relatedly, it has given rise to new concerns about resilience and redundancy in view of cyber-security, in an age of interconnectedness and reliance on communication and energy networks.

What, however, can be understood as 'algorithmic regulation'? Is there something clearly identifiable and distinct from other types of regulatory control systems that are based on standard-setting ('directors'), behaviour-modification ('effectors') and information-gathering ('detectors')?

One way to point to distinctiveness is that algorithms can 'learn' – and that the codes on which these algorithms are 'set' and 'learn' are far from transparent. A second component is the supposedly vast computing power in processing information. A third component is the large 'storage' capacity that potentially allows for comparison and new knowledge creation. A fourth component might be the insidious nature in which 'detection' does take place: users casually consent to highly complex 'conditions of service' and are not necessarily in control of the way in which their 'profile' is being processed. Similarly, and this is a fifth component, behaviour-modification is said to work by using architecture and 'nudges'. In other words, one might argue that algorithmic regulation is an *extension* to existing control systems in terms of their storage and processing capacity; they are qualitatively *different* in that much of the updating is performed by the algorithm itself (in ways that are non-transparent

to the external observer) rather than based on rule-based programming; and it is *distinct* in its reliance on observation and default-setting in terms of detecting and effecting behaviours.

At the same time, the notion of decision making and 'learning' by the algorithm itself is certainly problematic. No algorithm is 'unbiased' in that the initial default setting matters, and so does the type of information that is available for updating. To maintain 'neutral' algorithms might therefore require biased inputs so as to avoid highly undesirable and divisive outcomes. Instead, what is called here 'by the algorithm itself' is that the ways in which these algorithms 'learn' and what kind of information they process is not necessarily transparent, not even to those who initially established these codes. This means that understanding the 'predictions' of algorithms is inherently problematic; they resemble the multiple forecasting models used by hurricane watchers where one day's 'perfect prediction' might be completely 'off' the following day.

Beyond these general debates about algorithmic regulation, there are a number of critical issues for regulation. Firstly, what is the impact of algorithms on 'users'? To some extent, one might argue that algorithmic regulation brings in new opportunities for users – it generates powerful comparisons that potentially grant users greater choice options on the market (and quasi-market) place than before. Similarly, algorithmic regulation can also be said to increase the potential for 'voice': enhanced information can be used for a more powerful engagement with users (e.g. users of public services). The threat of 'choice' and 'voice' might make providers of services more responsive to users.

However, as individual experiences disappear into 'big data', engagement is mediated. This, in turn, points to the requirement that we need to understand better the ways in which user experiences are mediated – and through which means.[1] Different means of mediating such experiences exist – it might be based on explicit benchmarking and league-tabling (thereby relying on competitive pressures), or it might be based on providing differentiated analyses so as to facilitate argumentation and debate, or it might be based on enhanced hierarchical oversight. Furthermore, as noted, algorithms are not neutral. They are therefore not just mediation tools but are instead of a performative and constitutive nature, potentially enhancing rather than reducing power asymmetries. In short, the regulation by algorithm calls for the regulation of the algorithm in order to address their built-in biases.

Secondly, and relatedly, as regulation via algorithm requires regulation of the algorithm, questions arise as to what kind of controls are feasible. In debates about the powers of state surveillance (in the context of Snowden), one argument has been made that the state's 'intelligence' powers are more accountable than those of private corporations. Such a view is controversial, but it raises the question as to how state and non-state actors should be held

[1] The notion of 'google knowing' describes the phenomenon in which the top search results' content is adopted in unquestioned ways.

accountable (i.e. reporting standards potentially backed by sanctions) and transparent (i.e. allow for external scrutiny). Transparency might also increase potential vulnerability to manipulation. Given the transnational nature of much corporate activity, it raises also the question of jurisdiction and the potential effects of national and regional regulatory standards (such as those relating to privacy).

Even if such regulatory oversight powers could be established (among state, non-state or para-state bodies), the regulation of the algorithm might also be expected to give rise to a new kind of regulatory analyst. Arguably, this means that this is the age of the forensic data analyst and programmer rather than the lawyer and the economist. Altering regulatory capacities in that way may prove challenging in itself. However, it is also likely to be challenging as the analytical capacities of the 'forensic data analyst' need to be combined with other capacities in terms of delivery, coordination and oversight. It also requires new types of combinations of analytical capacities; for example, when it comes to the regulation of information, it is not just the presentation of particular 'facts' that requires monitoring, but it is increasingly their visualisation. In the field of energy, it requires the combination of engineering and data analysis.

Furthermore, there is the question at what point such regulation of the algorithm could and should take place. One central theme in ethical debates has been the default setting – algorithms should not be set to make straightforward ethical choices, but should be programmed so as to make 'context-dependent' choices. Such a perspective is problematic as no algorithm can be 'neutral'. As information can emerge and 'wiped' or deleted (but not everywhere), and as complex information systems generate new types of vulnerabilities, as information itself can be assessed in remote (non-intrusive) ways, regulatory capacity is required to deal with information in 'real' rather than 'reactive' time.

The third central issue for the regulation of algorithms is vulnerability to gaming and corruption. We define 'gaming' as the use of bots and other devices to mislead: information flows are generated that might, at first sight, appear as 'real', but, on second sight, reveal that they are generated by artificial means and/or are inflated so as to provide greater visibility to some 'information' than others. This might be related to the use of social media to communicate certain messages, or it might be used to enhance the visibility of certain websites on search engines. In contrast, corruption is the explicit attempt to undermine the functioning of the system rather than its exploitation. This is therefore the world of cyber-security and the protection of critical infrastructures (that increasingly operate in the cloud without sufficient protocols to deal with 'black swans', let alone, 'fancy (or cozy) bears' (Haba 2017).

In response, it might be argued that regulation *by* algorithm makes gaming also less likely when it comes to oversight. Performance management by target and indicator is widely said to suffer from extensive gaming and manipulation (i.e 'corruption'). The power of algorithms to deal with information could be said to enhance the possibilities of regulators to vet information in unpredictable ways,

thereby reducing the opportunities by organisations to game. However, as the work by Alex Griffiths and colleagues has shown, assessing complex organisations via algorithms remains a difficult undertaking that does not necessarily enhance the predictive powers of regulatory oversight.

Beyond these questions of the regulation of algorithmic regulation remains the wider concern with ethical questions. As has been demonstrated, artificial intelligence devices can quickly turn racist (Devlin 2017; Kleinman 2017) as they process embedded information and their explicit and implicit biases. It raises issues about the transboundary effects of national (state and non-state) efforts to set standards and it also raises issues about the differential interests of users – insisting on 'privacy' on the one hand, but also demanding 'ease of use' on the other. Finally, it also raises the ethical question about the nature of public policy: what kind of expertise should be prioritised? Table 1 summarises our argument about the potential effects of regulation by algorithm.

Table 1: Potential effects of regulation by algorithm

Increased contrived randomness + makes gaming and corruption less feasible as regulators can process vast information flows rather than rely on key indicators - Complex and vast information might reduce possibility of detecting essential information/non-transparency of the algorithms means lack of understanding of patterns	Increased oversight + makes risk-based assessments more likely as vast information flows allow for more fine-grained analysis and bespoke oversight - Substantially enhances intrusiveness and surveillance powers
Increased rivalry + enhances possibility for ranking and benchmarking - Enhances vulnerability to gaming and corruption by bots and malware attacks	Increased mutuality + enhances information for informed engagement - Increases dominance of 'data analyst' over other kinds of professional knowledge/biased conversation

In sum, therefore, the question of how to deal with the regulation of algorithms returns us to the underlying normative position established by Harald Laswell in his call for an interdisciplinary field of 'policy analysis', namely the need for a population with knowledge *of* and *in* the policy-making process.

References

Devlin, H. (2017) 'AI programs exhibit racial and gender biases, research reveals', *Guardian,* 13 April 2017, <https://www.theguardian.com/technology/2017/apr/13/ai-programs-exhibit-racist-and-sexist-biases-research-reveals> Accessed 18 September 2017.

Haba, M. (2017) 'Black swan in the cloud'*, risk&regulation* 33: 16–17. <http://www.lse.ac.uk/accounting/CARR/pdf/Risk&Regulation/r&r-33/riskandregulation-33-black-swan-in-the-cloud.pdf> Accessed 18 September 2017

Kleinman,Z. (2017) 'Artificial intelligence: How to avoid racist algorithms', BBC News, 14 April 2017 <http://www.bbc.co.uk/news/technology-39533308> Accessed 18 September 2017.

Martin Lodge and Andrea Mennicken, Centre for Analysis of Risk and Regulation, London School of Economics and Political Science.

Algorithms, governance and regulation: beyond 'the necessary hashtags'

Leighton Andrews

The necessary hashtags

What do you call one thousand lawyers replaced by robots? In this variant of the old joke, the answer remains 'a good start'. But I am writing on the day that the Serious Fraud Office has revealed that it recently used algorithmic software to review '30 million documents at a rate of up to 600,000 a day, whereas a team of barristers would previously have processed 3,000' (Bridge 2017).

At the end of March 2017, the UK Home Secretary was roundly mocked after she referred to the need to call on the support of 'the best people who understand the technology, who understand the necessary hashtags' (Mezzofiore 2017), to take action against terrorist messaging and posting on social media and messaging platforms. Analysing politicians' language may not be the best route to evaluating the readiness of governments to address complex issues such as the regulation of algorithms, artificial intelligence and machine learning, but it is one indicator of the challenges that face those seeking to advance political understanding and build a platform for action. It will be hard to build public confidence if senior policymakers are not seen as credible explainers of the challenges. Other politicians may have deeper understanding. Angela Merkel was very specific when she said of Facebook and Google in October 2016:

> The big internet platforms, via their algorithms, have become an eye of a needle, which diverse media must pass through to reach users … These algorithms, when they are not transparent, can lead to a distortion of our perception, they narrow our breadth of information (BBC News 2016).

Meanwhile, President Obama was comfortable enough to guest-edit *Wired* magazine and to explain why 'government will never run the way Silicon Valley runs' (White House 2016), at a technology 'Frontiers' event last autumn – but the digital challenges now facing governments go well beyond the development of citizen-friendly services on public digital networks. A growing range of challenges driven by advances in artificial intelligence and machine learning are going to require an expanded digital confidence and capacity from politicians and regulators. 'Algorithmic accountability' doesn't easily lend itself to a manifesto pledge, but across a range of sectors it is becoming an increasingly important issue – and it is not clear that the political, administrative or regulatory capacity has evolved to address the challenges.

Algorithmic harms

Perhaps it is helpful to consider at the outset what are the dangers against which we are trying to protect ourselves? Why have algorithms come into such public prominence over recent years? First, we have well documented examples of *algorithmic bias*, in which judgements on individual futures – employment, eligibility for loans, likelihood of imprisonment – are determined by algorithmic choices which have in-built human errors or conscious or unconscious biases. Second, we have clear examples of *algorithmic manipulation*, in which judgements about, for example, news, information or advertising, are constructed on the basis of data collected on individuals and used to channel what is presented according to inferred preferences. Third, we have perceived or actual *algorithmic lawbreaking*, in which algorithms are apparently deliberately constructed to deceive lawmakers and regulators, for example, in terms of emissions controls or traffic management, or attempts at price-fixing. Fourth, we have growing evidence of algorithm usage in *propaganda*, from disinformation campaigns by unfriendly countries to election campaign bots. Fifth, there is the issue of *algorithmic brand contamination and advertising fraud* where major brands have found their advertising placed alongside hate speech or terrorist material, or where human interaction with the advertising is proven to be less than reported as bots are upping the claimed strike-rate. Sixth, there is what I call *algorithmic unknowns* – the question of how machine learning means algorithms are becoming too complicated for humans to understand or unpick; it is in this arena also that the concern about general AI taking anthropomorphic forms and following the long-standing themes of 'technics-out-of-control' also emerges (Winner 1977).

I don't pretend that is a comprehensive list, but it will suffice as an illustration of the challenges that now face and will face regulators across a range of fields.

Regulating the algorithm

I was first involved in attempts to regulate algorithms over 20 years ago. Working at the BBC, we sought to regulate the new gatekeeping technologies of digital television – what were commonly known as 'set-top boxes' – driven by conditional access systems whose algorithms determined what content viewers could access, depending on the subscriptions they had paid (see Levy 1997). These were simple algorithms – and we were actually seeking to regulate corporate behaviours, rather than the algorithm itself. Today, it seems that there are three kinds of behaviour which call into question the need for regulation, crudely summarised as:

- Human behaviour (initial encoding of, development of training data for and management of algorithms)
- Corporate behaviour (proprietary algorithms and their deployment, management and governance)

- Algorithmic behaviour (the black box issue, machine learning)

Andrew Tutt (2016, see Table 1 below) has suggested a qualitative scale of algorithmic complexity which may be helpful in assessing the nature of risk to society and how that might be managed or regulated:

Table 1: A possible qualitative scale of algorithmic complexity (Tutt 2016: 107)

Algorithm type	Nickname	Description
Type 0	'White box'	Algorithm is entirely deterministic (i.e. the algorithm is merely a pre-determined set of instructions.
Type 1	'Grey box'	Algorithm is non-deterministic, but its non-deterministic characteristics are easily predicted and explained.
Type 2	'Black box'	Algorithm exhibits emergent proprieties, making it difficult or impossible to preduct or explain its characteristics.
Type 3	'Sentient'	Algorithm can pass a Turing Test (i.e. has reached or exceeded human intelligence).
Type 4	'Singularity'	Algorithm is capable of recurive self-improvement (i.e. the algorithm has reached the 'singularity')

It is important to stress, lest, in the light of public attention that has recently been given to use of algorithms by Google, Facebook, Volkswagen and Uber in particular, algorithms are thought to be ungovernable, that regulatory authorities do have some experience of regulating algorithms. The Financial Conduct Authority (FCA), to take but one example, in the field of high frequency trading (HFT; see FCA 2014).[1] Regulation is difficult, and never perfect, but possible.

What are the policy instruments under consideration?

The recent proposed inquiry by the House of Commons Science and Technology Committee (2017), into 'Algorithms in decision-making', suspended by the General Election, drew a significant amount of evidence. A simply policy instrument analysis of the evidence submitted shows the following proposals (see Table 2).

So a range of technical, governance, regulatory, legislative or institutional proposals were outlined. Notable was the absence of any serious *fiscal* proposal. Meanwhile Bill Gates (Waters 2017) has called for the taxation of robots in order to replace the payment of income tax lost to automation, though as Floridi (2017) has pointed out it is not clear how this might be done. First there would be the issue or how to define a robot; second, given that certain kinds of robotic device have been in use in automated factories for some time, the question of retrospection would also arise; third, it is not *yet* clear that automation will necessarily lead to a new displacement of jobs. Income tax may be a policy blind alley.

[1] See also FCA's 'Content of proposed MAR 5 & MAR 5A – systems and controls for algorithmic trading on MTFs and OTFs'. <https://www.fca.org.uk/mifid-ii/8-algorithmic-and-high-frequency-trading-hft-requirements>

Table 2: Policy instrument analysis of the House of Commons Science and Technology Committee evidence on algorithms in decision making

Technical	Governance	Regulation	Legislative	Institutional	Fiscal
• Transparency, accountability and explicability • Best practice • Training data of algorithm to be prescribed • Distinctions between basic and machine learning algorithms • Further research on technical mechanisms to interrogate 'black box'	• Internal compliance teams • GDPR compliance certification • Public sector algorithms to be analysed in line with MacPherson review of government modelling • Develop professional standards for data science • Support role of Partnership for AI	• Sectoral statutory oversight body (e.g. police) • New scrutiny and oversight duties on existing regulators • Requirements not to design algorithms which challenge protected characteristics under HR law • Requirement for EIAs or HRIAs for algorithms • GDPR as basis of regulation • Medical algorithms to be subject to MHRA • CMA to investigate pricing algorithms	• Legally mandated 'right to explanation' of automated decisions to supplement GDPR • Humanly interpretable decision-making methods in mandated risky sectors • Categorisation of risky and non-risky sectors/mechanisms • Right to challenge by those affected • Mandation of certification mechanisms to ensure fair, open and non-discriminatory practices prior to deployment of algorithms	• Algorithmic or Machine Learning or Data Ethics oversight institution with proper resourcing OR/AND capacity building for existing regulators • Investment in public R&D on algorithms • Technical oversight and template design • Analyse algorithmic experience from credit scoring industry as potential for best practice in accountability	• None

However, there are other ways of using fiscal instruments where in the past, these have been used as ways of mitigating risk: the variable charges on car tax dependent on their emissions of noxious gases, for example. Radio and television licences have been means of funding developments in new technologies (for example, the colour TV licence as an additional licence fee payment in the 1970s). Public goods such as radio-communications spectrum and independent terrestrial television franchises have been subject to auction. Is there a case for creating new fiscal instruments for algorithms, artificial intelligence and machine learning? For example, to limit insecure devices connected to the Internet of things, should there be a *connectivity tax*? Or to as insurance against harms by 'home-companion' robots, or for first generation driverless vehicles, should these come on-stream, or to raise additional funds from algorithmically-driven Internet intermediaries to fund other media?

Conclusion: an oversight institution?

There have been a variety of proposals for some kind of oversight institution: an AI Watchdog (Sample 2017), a Machine Learning Commission (Mulgan 2016), or in the US context, an FDA for Algorithms (Tutt 2016), or a National Algorithm Safety Board (Macaulay 2017). The recent Royal Society and British Academy report (2017) on Data Governance has reinforced this. A single regulatory or ethics body may be appropriate in each jurisdiction, though there will be a need

for international coordination. Additionally, there will be sector-specific challenges on algorithmic regulation, and a single AI watchdog is unlikely of itself to have the time or capacity address *all* issues, and could be a heavy-handed instrument, or worse, a delay to innovation. It may be part of the solution, as an umbrella, over-arching body with supervisory and advisory roles, but not the whole solution as sectoral regulators need to strengthen their capacity in this area. Indeed, as I said in written evidence to the House of Commons Culture, Media and Sport inquiry into Fake News, some of these issues raise questions which cross the boundaries of sectoral regulators (Andrews 2017).

We have a new government with a manifesto commitment to creating an expert Data Use and Ethics Commission 'to advise regulators and parliament'; a regulatory framework for data and the digital economy; 'a sanctions regime to ensure compliance' and a power for an industry-wide levy on social media providers and communications services providers to support awareness and preventative activity. So we are, it seems, moving beyond 'the necessary hashtags'.

References

Andrews, L. (2017) 'Fake News and the threat to real news', Written evidence to the House of Commons Culture, Media and Sport Select Committee, <http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/culture-media-and-sport-committee/fake-news/written/48139.pdf>

BBC News (2016) 'Angela Merkel wants Facebook and Google's secrets revealed', 28 October. <http://www.bbc.co.uk/news/technology-37798762> Accessed 18 September 2017.

Bridge, M. (2017) 'Serious Fraud Office calls in robot to solve cases such as Rolls-Royce corruption', *Times*, 27 June. Accessed 18 September 2017.

British Academy and the Royal Society (2017) 'Data management and use: governance in the 21st century'. Joint report, June 2017. <https://royalsociety.org/~/media/policy/projects/data-governance/data-management-governance.pdf> Accessed 18 September 2017.

Financial Conduct Authority (FCA) (2014) 'Regulating high frequency trading', 4 June, <https://www.fca.org.uk/news/speeches/regulating-high-frequency-trading> and 'Content of proposed MAR 5 & MAR 5A – systems and controls for algorithmic trading on MTFs and OTFs'. Accessed 18 September 2017.
<https://www.fca.org.uk/mifid-ii/8-algorithmic-and-high-frequency-trading-hft-requirements>

Floridi, F. (2017) 'Robots, jobs, taxes and responsibilities', *Philosophy and Technology* 30 (1): 1–4. <https://link.springer.com/article/10.1007/s13347-017-0257-3>

Levy, D.A.L. (1997) 'The regulation of digital conditional access systems. A case study in European policy making', *Telecommunications Policy* 21(7): 661–76.

Macaulay, T. (2017) 'Pioneering computer scientist calls for National Algorithm Safety Board', Techworld, 31 May. <http://www.techworld.com/data/pioneering-computer-scientist-calls-for-national-algorithms-safety-board-3659664/> Accessed 18 September 2017.

Mezzofiore, G. (2017) 'Politician's baffling quote about hashtags gets the mocking it deserves', MashableUK, 27 March. <http://mashable.com/2017/03/27/necessary-hashtags-whatsapp-encryption/#7Agb2Is1Umq2> Accessed 18 September 2017.

Mulgan, G. (2016) 'A machine intelligence commission for the UK: how to grow informed public trust and maximise the positive impact of smart machines', Nesta, February. <http://www.nesta.org.uk/sites/default/files/a_machine_intelligence_commission_for_the_uk_-_geoff_mulgan.pdf> Accessed 18 September 2017.

Sample, I. (2017) 'AI watchdog needed to regulate automated decision-making, say experts', *Guardian*, 27 January, <https://www.theguardian.com/technology/2017/jan/27/ai-artificial-intelligence-watchdog-needed-to-prevent-discriminatory-automated-decisions> Accessed 18 September 2017.

Science and Technology Committee, House of Commons (2017) 'Algorithms in decision-making inquiry'. <https://www.parliament.uk/business/committees/committees-a-z/commons-select/science-and-technology-committee/inquiries/parliament-2015/inquiry9/>

Tutt, A. 2016. 'An FDA for algorithms?' *Administrative Law Review* 69: 83–123. <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2747994>

Waters, R. 2017. 'Bill Gates calls for income tax on robots,' *Financial Times*, 19 February. Accessed 18 September 2017.

White House (2016) 'Remarks by the President in Opening Remarks and Panel Discussion at White House Frontiers Conference', 13 October. <https://obamawhitehouse.archives.gov/the-press-office/2016/10/13/remarks-president-opening-remarks-and-panel-discussion-white-house> Accessed 18 September 2017.

Winner, L. (1977) *Autonomous technology*, Cambridge MA: MIT Press.

Leighton Andrews, Cardiff Business School, Cardiff University, Wales.

Evaluating predictive algorithms

David Demortain and Bilel Benbouzid

Algorithms and models for automated analysis of data are developed with a great deal of promises in mind, about precision, completeness, up-to-datedness, optimisation and anticipatory capacity gained by organisations employing these technologies. Many of these promises and expectations will, of course, be denied in practice, and algorithms may even produce consequences in decision that are adverse to the values which are meant to drive the regulatory decision in the first place.

One first possible form of regulation applying to algorithms, thus, could be some kind of product-based regulation of quality, applying quality standards to these (software) products, just like regulatory regimes were invented over time for other products and technologies through the course of innovation. Standards could thus define what algorithms or the IT systems, by which they are deployed, should have. Firstly, one would think they should have a certain level of transparency about what's in the system: the core algorithm itself, or the code. A second approach could be to adopt the kind of instruments that has generalised under the generic regime of 'risk regulation', in particular, those that that seek to evaluate or assess, *ex ante* as well as *ex post*, the effects associated with a product or technology, including not very probable effects.

The first form of 'product'-based regulation may perhaps be considered more appropriate, in the light of problems of the opacity and the 'black-boxed' nature of algorithms, and general difficulties in accessing the heart of the IT systems and their codes. But the second form of regulation may be at once more difficult and more likely. More difficult, first, because the standards or benchmarks against which one judges that an algorithm produces a systematic effect, and that this effect is negative or adverse, are generally not easy to establish; there is a basic problem in knowing that an algorithm produces an effect on something that we generally cannot know otherwise. Algorithms, by definition, are systems that produce knowledge or treat data at scales and levels of complexity that no other body of knowledge or experience can access. There may be experience available to judge that an algorithm gets things wrong, but then the question becomes the availability of this experience, its codification, and how open and honest the process of evaluating or validating the algorithm against this experience will be. More likely, however, because the structure of the industry in which algorithms are developed, and the competition within that industry between commercial and academic or public developers may be, in some cases, conducive to the emergence of forms of testing and comparative

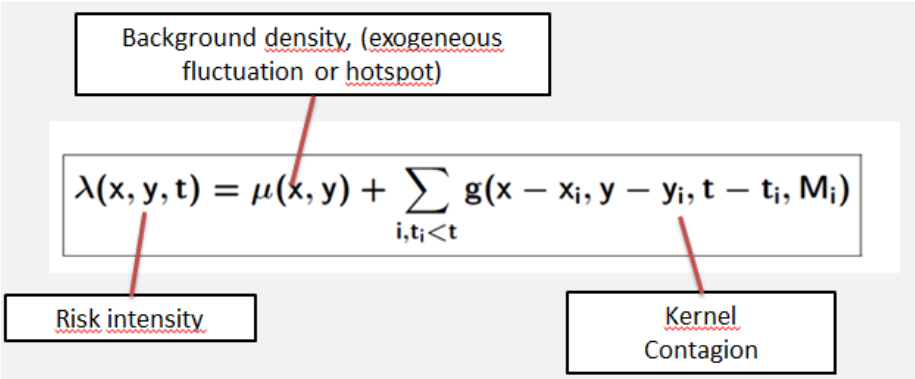
trials of algorithms, on which more formal regulatory systems could build. I draw from the area of predictive policing to illustrate these issues.

Crime analysis tools have proliferated in the last decades,[1] the more innovative or recent development being crime series analysis tools, that are also termed 'predictive policing' tools – a label that many reject, preferring instead to stick to a more modest phrase of crime analysis.

Among those tools that aim to predict crime, PredPol has received the most attention, because of its claims to be able to predict crime, and its commercial success too. PredPol is a software that claims to be able to do more than spatial analysis and identification of 'hot spots', to extrapolate from existing data, with precision, territories to patrol to avert crime. PredPol as a company does not easily lend access to its algorithm or to the data on which it is used, and thus does not help 'audit' or validate the predictions that it makes. It is well known, however, that they took inspiration from an algorithm developed for the purpose of predicting earthquake aftershocks. The research that influenced PredPol is that of David Marsan, a professor in the earth science laboratory at the University of Savoie, in Chambéry, France, and a specialist in the study of earthquake aftershocks.

The algorithm that he developed was introduced into the PredPol system, on the basis of the principle that crimes are subject to 'after-crimes' in the same way as earthquakes are followed by aftershocks. In other words, PredPol imported a geo-physical theory of 'loading' of earthquake potential, into the analysis of crime, through the notion of 'contagion'. Crime occurrence can be predicted by jointly calculating hotspots, and a potential of contagion from one crime to the next. The particularity of this method is that it is a very 'lean' model, with a minimal number of parameters in the equation. The quality of the predictions depends on the theory of contagion, rather than on the completeness of parameters and of the data entered into the system.

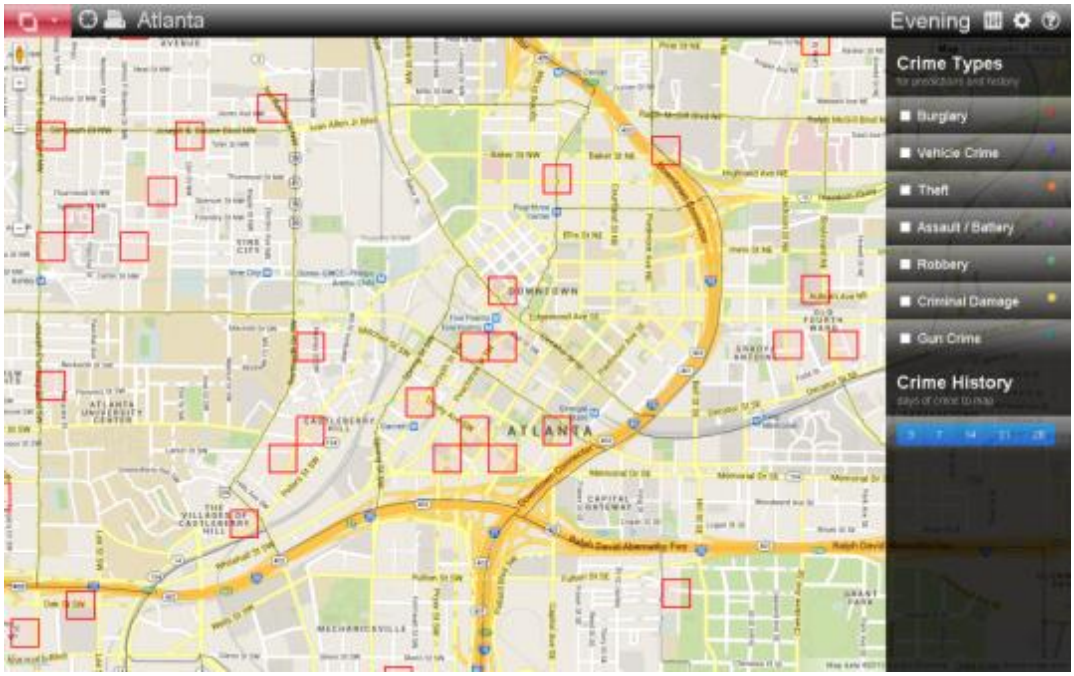
Figure 1: Predicting occurrence



[1] *Smart Policing Initiative Quarterly*, Newsletter no. 7, Springer 2013.

So PredPol generates maps of the territories to patrol, algorithmically governing the work of police officers in the field, in the name of scientific accuracy. In many American cities that have purchased the tool, police officers can check these predictions directly on a tablet, or work with print-outs of the maps with 150m x 150m cells highlighted.

Figure 2: Mapping territories



It is by definition nearly impossible to 'validate' the predictions made by the tool or assess their impact on crime minimisation. Predictions are nearly unfalsifiable. On the one hand, they are very unlikely to bring police officers to the site of a crime, and help them catch the offender red-handed. At the same time, this does not discourage PredPol from arguing that their tool is predictive. The argument they use is that, quite simply, the presence of a police patrol on a territory will discourage crime offenders. This is hard to prove too. So the predictions by the tool are difficult to test and verify, as is the overall marketing claim of PredPol. An evaluation could be made of the quality of predictions, e.g. by collecting feedback from police officers about their experience with the tool, and their knowledge of the evolution of the territories in question and crime patterns, as they observe it. Or a correlation between the use of PredPol predictions and the evolution of crime in a set of cities, in which it is in use, could be computed. In both instances, PredPol seems to discourage this kind of evaluation, by protecting its core and lean algorithm as a trade secret.

An evaluation of the predictions was made, as a sort of sociological experiment by liaising with David Marsan, whose work on earthquake predictions PredPol took its inspiration. Marsan reconstituted the model that PredPol uses, and downloaded crime data for the city of Chicago, to generate predictions and

compare them with those of PredPol (that had released predictions for the same city in a 2014 publication). The conclusion of this informal validation exercise was that:

> These results cast strong doubts on the capacity of the models proposed here to outperform simple hotspot maps obtained by smoothing, for the dataset analysed. The triggering contribution to the occurrence of future events is small (it accounts only for 1.7 % for the best model). Accounting for memory in the system therefore can only provide a very modest contribution to the effectiveness of the prediction scheme. […] More importantly, it is assumed that the dynamics of the process stays the same over time. Possible non-stationarity of the process is thus clearly an issue, as it will prevent the use of past information to predict the future.[2]

Marsan identified problems with the theory of contagion and its contribution to the prediction of crime as opposed to other parameters, but also to the assumption of stability/stationarity in crime events over time and finally, with the fact that predictions are performative as they trigger decisions that will affect the environment being predicted. For example, burglars respond to changes in policing strategies induced by predictions. So, 'unlike natural processes like earthquakes, analyses like the one presented here could therefore have the ability to modify the observed process, making it more difficult to, correctly predict future events'.[3]

The performances of PredPol have got to be understood better, but only in this informal manner unless or until more information is provided. Other ways of evaluating algorithms and their performances involve cross-city exchanges of experience, through more or less formalised contacts between police forces of various cities that use the commercial service, mainly after dissatisfaction with the service. The more critical development in terms of public evaluation of algorithmic performance came through the development of competitors of Predpol. HunchLab's own predictive system contrasts with PredPol in many ways. Firstly, HunchLab was established by a not-for-profit startup in data sciences, that has developed a range of applications in collaboration with police forces and academics. Secondly, it has moderate claims about its core algorithm and its pure predictive performance. As its developers argue, prediction is a loaded term that deflects from the other part of the phrase predictive policing. While they are confident in the quality of their predictions, they emphasise that their tool is one that helps design patrol allocations to minimise preventable, foreseeable harm. Thirdly, and relatedly, the algorithm itself is not conceived of as a major trade secret. Very little time has been spent on developing it. Much more time, however, was spent on defining the complex set of data and

[2] Pers. comm. David Marsan to Bilel Benbouzid, March 2015.
[3] Ibid.

parameters to model the environment of policing and crime, so as to best predict outcomes within it. They model many more aspects than PredPol, including environmental risk factors (presence of bars, prostitution in the area), but also weather and moon phases. Lastly, the performance of the prediction is not reduced to the validity of the predictions made by the algorithm, as verified *ex post* against actual crime offences measured in the area. The algorithm is judged 'performant' if and when officers actually trust and use it. To reach this level of 'trust' in the prediction, HunchLab incorporates in its tool various applications for feedback by the police officer, incorporating his or her experience in the prediction.

The development of HunchLab is in and of itself the result of an evaluation of the performance of the competing tool of PredPol, against which it is positioned in the market in multiple ways (commercial/non-commercial; predictive/decision-aiding tool etc.). The point here, again, is not whether an exercise of testing or validating the algorithm and its predictions actually occurs, or if an algorithm is judged wrong or right. The more intriguing point is how common evaluative knowledge about what the algorithm does emerges. There are many barriers to the production of this evaluation of algorithms, not least when algorithms are developed commercially, as part of proprietary products and services. This makes it apparently more difficult to produce the iterative tests that are necessary to arrive at an objective view of the performance of the prediction, and remaining uncertainties such as testing it on new territories, or more and varied sets of data, testing it *ex post* against actual crime occurrence, and so on. The various informal, indirect ways in which evaluative knowledge has emerged in the above case shows that there is potentially a gap here, that is only filled when someone admits the public relevance and importance of what the algorithm does, and takes up an arguably public role of testing its effects.

Other cases reveal the importance of the industry developing the algorithms, and of developers with a greater level of 'publicness', in the sense of alignment on public values and goals. The algorithms developed to quantitatively predict the risk of a toxic chemical, sold to chemicals-regulating agencies and businesses regulated by these in the 1990s, only slowly came to be criticised, and finally little used, as a small set of academics or scientists working in publicly funded health and environment research institutes took on the task of systematically investigating the proposed tools, and run them on a variety of domains, to produce an estimate of the quality of the prediction and the remaining uncertainties. As algorithms were found to be able to reproduce experiments in some 20 to 60 per cent of the cases, the claims of their developers concerning their accuracy, predictive capacity and utility for regulatory decision makers have gradually eroded. In this case, the existence of a public industry and profession of toxicological modelling allowed this to happen sooner than in the predictive policing area, where most developments are delegated to an emerging data science industry. Both cases, comparatively,

pose the question of the institutional structure enabling the evaluation of the knowledge offered by algorithms.

David Demortain and Bilel Benzouid, Laboratoire Interdisciplinaire Sciences Innovations Sociétés, Paris, France.

The practical challenges of regulating the quality of public services with algorithms

Alex Griffiths

This paper details some of the real-world challenges of implementing algorithmic approaches to regulating the quality of public services. Using the health and higher education sectors as case studies, it demonstrates the differences between theory and practice, followed by a short outline of the conditions necessary for algorithmic approaches to succeed, and a short conclusion which may form the basis of further discussion.

Risk-based regulation and the use of algorithms

Risk-based regulation is built upon the allocation of regulatory resource in proportion to the risks posed to the regulator's objectives (Black 2005; Rothstein et al. 2006b). It ostensibly provides practitioners with a means with which to 'maximise the benefits of regulation while minimising the burdens on regulatees by offering "targeted" and "proportionate" interventions' (Rothstein et al. 2006a: 97).

For public services, this means the ability to lighten or eliminate inspections for low-risk providers, leaving them free to prosper, whilst using the resource saved to conduct inspections of high-risk providers and quickly eliminate or prevent any poor practice, all at the same or reduced cost to the taxpayer. In theory, everybody wins.

Prioritising regulatory resource is, of course, not a new challenge (Pontell 1978). Risk-based approaches are intended to replace the implicit prioritisation of resource previously conducted behind closed doors with the explicit determination of risk through assessment frameworks (Black 2005: 4).

Regulators rarely publish comprehensive details on how they calculate risk; however, it is possible to loosely place their calculative approaches into three categories:

a. Typically utilised by regulators developing their first risk-assessment approach, regulatees can be assigned one of a small number of risk categories by means of a simple, and often contextual, *rules-based* assessment.

b. *Data-informed* prioritisation tools use algorithms to weight and aggregate an often large number of metrics selected a priori by experts to generate a risk rating and or report.

c. Data-driven approaches make use of machine-learning techniques to identify, weight and aggregate useful metrics and develop optimal statistical models without human interference.

The risk ratings generated by these algorithms may either be reviewed by humans who make the final prioritisation decision, or may automatically prioritise activity. Real-world examples of *data-informed* and *data-driven* algorithmic approaches are detailed below.

Algorithms and the Care Quality Commission (CQC)

The CQC is responsible for regulating the quality of care provided at all 30,000 health and social care organisations in England. Regulating such a large number of organisations providing often vital and/or dangerous services to vulnerable users has resulted in CQC fully embracing a risk-based approach.

In the NHS, there is a wealth of data including: waiting times, mortality and readmission rates, staff surveys, patient surveys, patient-led estates assessments, infection rates, finance and governance measures, staff qualifications, staffing levels and hours worked, whistleblowing reports, and safety notifications.

The complex 'Quality and Risk Profiles' (QRPs) were developed with expert statistical input to aggregate approximately 1,000 indicators mapped to one of 16 care 'outcomes'. Each indicator had three individual weightings and a score on a seven-point scale, and these weighted scores were in turn aggregated to generate a risk score for each of the 16 'outcomes' on an eight-point scale. Each QRP was updated nine times a year.

Despite the wealth of data and (relatively) complex algorithm, QRPs were not well regarded by inspectors and, more importantly, failed to identify risks to the quality of care (Walshe and Phipps 2013). The median risk score preceding an inspection findings of 'minor non-compliance' was actually lower than the median risk score preceding an inspection finding of compliance.

Following criticism in the Francis Inquiry into the scandal at Mid-Staffordshire NHS Foundation Trust, CQC made significant changes including its risk tool. 'Intelligent Monitoring' was designed to be far simpler and, following consultation with the sector, 150 indicators which it was felt would best identify risks to the quality of care were selected for the tool (CQC 2013). Each indicator was categorically scored as either 'No evidence of risk', 'Risk', or 'Elevated risk' and aggregated with equal weight (CQC 2014).

The revised and far simpler 'Intelligent Monitoring' tool was, however, also unable to successfully identify risks to the quality of care. Indeed, the regular scoring and aggregation of 150 expertly chosen indicators was actually wrong

more often than it was right. It would have been marginally better for the CQC to do the exact opposite of what the tool suggested (Griffiths et al. 2016).

Therefore, even with arguably more data than any other regulator in the UK, and two quite different approaches, CQC was unable to successfully automate the collection and scoring of data to prioritise its activity. This, however, does not necessarily mean it cannot be done. It may be the case that CQC have not yet found the right algorithm – something which could be determined via machine learning. This approach was adopted by the Quality Assurance Agency.

Algorithms and the Quality Assurance Agency (QAA)

The QAA are responsible for assuring quality standards in UK higher education. Later than others to adopt the risk-based approach, the 2011 Higher Education White Paper 'Students at the Heart of the System' called for QAA to adopt:

> … genuinely risk-based approach, focusing QAA effort where it will have most impact and [to] explore options in which the frequency – and perhaps need – for a full, scheduled institutional review will depend on an objective assessment of a basket of data, monitored continually but at arm's length (BIS 2011: 3.19)

Unlike the development of CQC's tools, a machine-learning approach was used to devise the optimal model for QAA using thousands of indicators including confidential QAA not available to the public and 600 sets of financial accounts purchased from Companies House. Despite having the outcome of all QAA inspections, an extremely comprehensive data set covering ten years and machine-learning techniques, no model could be developed that successfully predicted the outcome of QAA reviews. Put simply, there was no relation between the available data and the outcome of QAA reviews. Even with perfect hindsight, the reviews could not be successfully prioritised based on the data (Griffiths 2017).

Why can't the data drive effective regulation?

It has therefore been empirically demonstrated that algorithmic regulation is not guaranteed to be successful. Whilst it is not possible to identify the specific reasons for the failure in the algorithmic regulation, it is possible via the above investigations and further work carried out as part of a recent King's College London project to identify potential contributing factors:

a. Data may be poor quality due to human error, poor information systems of varying degrees of 'gaming' incentivised by the growing use of indicators in performance management.
b. Data may be correct but of limited use by the time it can be processed and acted upon by the regulator due to its age. This is especially true of annual data collections.

c. Data may be being misused, for example student satisfaction survey results are used as a proxy for teaching quality in spite of the fact students may be most happy not being challenged and guaranteed a high grade.
d. Data may be too coarse, capturing data at university- or trust-level may average out any signs of variation across the large and complex organisations and fail to pinpoint pockets of poor quality.
e. Inspections findings are inherently constrained by what inspectors can see and understand in large, specialist organisations in a short period of time.
f. The outcome of inspections are often over-reductionist, for example describing the quality of a multi-billion pound, multi-hospital NHS trust with a single word.
g. Processes being assessed by regulators may have become entirely decoupled from the outcomes assessed by data as regulatees are keen to demonstrate compliance and pass their 'ritual of verification' (Power 1997).
h. Data and inspections may simply be assessing different things.
i. Algorithmic techniques work best when there is a large number of data-rich cases so that statistical associations can be found between inspection findings and the data on which the probability of those inspection findings statistically depend; predictive power may be inherently constrained by a limited number of regulatees and inspection findings.
j. Algorithms will struggle to predict the outcome of quality inspections when the nature of 'quality' is ambiguous, unstable and contested; for example, in higher education should university quality be assessed in terms of student employability, satisfaction, retention, widening participation or A-level tariffs? Even if quality is clearly defined, whether an inspector feels it is being provided is inherently subjective in fields such as health and higher education.

Conclusion

If algorithmic regulation can be successfully achieved it offers many benefits with few if any drawbacks. Its appeal to government is clear. It ostensibly allows them to: reduce the amount of resource they have to invest in regulators showing themselves as efficient with taxpayers' money, be seen to be embracing new technologies, prevent or quickly respond to poor quality, and actively cut the 'red tape' that is burdens good providers.
For algorithmic regulation to be a success however the real-world problems which limit its application must be acknowledged and overcome. To ignore these challenges is easy for politicians as successful algorithmic regulation comes with all the above benefits and any failings can be blamed on poor implementation by the regulator. Moreover, anyone stating it can't be done runs the very real risk of appearing monolithic and open to criticism that the problem is not with algorithmic approaches, but with their inability to implement it. Until these

problems are acknowledged and addressed, algorithmic regulation of the quality of public services will continue to fail and it is the people that regulation was designed to protect who will suffer.

It is tempting, as QAA have done, to suggest any problems with algorithmic regulation can be sorted by having a panel of experts review the output of any risk model and make the final decision (Kimber 2015). This can have its advantages; higher education specialists know that low contact hours at the University of Oxford result from their individual tuition approach rather than neglecting students and can compensate for this. However, humans are prone to numerous heuristics and biases which mean these correct interventions are outweighed by incorrect interventions (Kahneman 2011).

It has been assiduously demonstrated that experts are, at best, equally as good (or bad) at making decisions as simple models (see for example Grove et al. 2000). Even when experts have access to the output of simple models, they have still been shown to perform worse than the model by itself (see for example Goldberg 1968; Montier, 2009).

A bad model is not made good by humans interpreting it. As the data and technology landscape changes, new opportunities arise which may solve some of the problems highlighted in this paper. For example, machines can – with significant effort – be taught to classify millions of items of student or patient feedback which cannot realistically be read or consistently coded by humans. Further, student location and engagement with online learning environment can be monitored to give new insight into student engagement with universities.

References

Black, J. (2005) 'The emergence of risk-based regulation and the new public risk management in the United Kingdom', *Public Law* 3**:** 512–48.

BIS (2011). 'Higher education: students at the heart of the system', White Paper, Cm 8122. London: Department for Business Innovation & Skills.

CQC (2013) 'Proposed model for intelligent monitoring and expert judgement in acute NHS Trusts (annex to the consultation: changes to the way CQC regulates, inspects and monitors care services'. Newcastle upon Tyne: Care Quality Commission. <http://www.cqc.org.uk/sites/default/files/documents/cqc_consultationannex_2013_tagged.pdf>

CQC (2014) 'NHS acute hospitals: indicators and methodology guidance to support the December 2014 Intelligent Monitoring update'. Newcastle upon Tyne: Care Quality Commission. <http://www.cqc.org.uk/sites/default/files/20141127_intelligent_monitoring_indicators_methodology_v4.pdf> Accessed 18 September 2017.

Goldberg, L.R. (1968) 'Simple models or simple processes? Some research on clinical judgments', *American Psychologist* 23(7)**:** 483–96.

Griffiths, A. (2017) 'Forecasting failure: assessing risks to quality assurance in higher education using machine learning', PhD thesis, King's College London. <https://kclpure.kcl.ac.uk/portal/files/67117431/2017_Griffiths_Alexander_1024268_ethesis.pdf>

Griffiths, A., Beaussier, A.-L., Demeritt, D. and Rothstein, H. (2016) 'Intelligent Monitoring? Assessing the ability of the Care Quality Commission's statistical surveillance tool to predict quality and prioritise NHS hospital inspections', *BMJ Quality & Safety,* <http://qualitysafety.bmj.com/content/26/2/120>

Grove, W.M., Zald, D. H., Lebow, B.S., Snitz, B.E. and Nelson, C. (2000) 'Clinical versus mechanical prediction: a meta-analysis', *Psychological Assessment* 12(1)**:** 19–30.

Kahneman, D. (2011) *Thinking, fast and slow*, New York: Farrar, Strauss and Giroux.

Kimber, I. (2015) 'Metrics and quality: do the numbers add up?' <http://wonkhe.com/blogs/metrics-and-quality-do-the-numbers-add-up/> Accessed 18 September 2017.

Montier, J. (2009) *Behavioural investing: a practitioners guide to applying behavioural finance*, Chichester: John Wiley & Sons.

Pontell, H.N. (1978) 'Deterrence theory versus practice', *Criminology* 16(1)**:** 3–22.

Power, M. (1997) *The audit society: rituals of verification,* Oxford: Oxford University Press.

Rothstein, H., Huber, M. and Gaskell, G. (2006a) 'A theory of risk colonization: the spiralling regulatory logics of societal and institutional risk', *Economy and Society* 35(1)**:** 91–112.

Rothstein, H., Irving, P., Walden, T. and Yearsley, R. (2006b) 'The risks of risk-based regulation: insights from the environmental policy domain', *Environment International* 32(8)**:** 1056–65.

Walshe, K. and Phipps, D. (2013) 'Developing a strategic framework to guide the Care Quality Commission's programme of evaluation', Report commissioned by CQC, Manchester: Manchester Business School, University of Manchester.

Alex Griffiths, Centre for Analysis of Risk and Regulation, London School of Economics and Political Science

Algorithmic regulation on trial? Professional judgement and the authorisation of algorithmic decision making

Jeremy Brice

Delegating regulatory decision making to algorithms is, we are often warned, a high stakes business. Regulatory decisions such as the determination of risk scores or the authorisation of investigation and enforcement action can have far reaching consequences for those affected by the courses of action (or inaction) that they set in motion. As such, it will be vitally important to ensure that algorithms are capable of making 'sound' decisions before they are introduced into regulatory processes – much as the judgement of human beings is typically validated through examination and accreditation before they are empowered to act as authorised officers of regulatory bodies. But what counts as sound judgement in the field of algorithmic regulation? How is one to tell whether an algorithm is 'performing well,' or making the 'right' decisions?

I was recently invited to tackle these questions when a middle ranking civil servant named 'Daniel'[1] approached me at the desk in the headquarters of the UK Food Standards Agency (FSA) where I work for three days a week as an embedded researcher. Daniel, who is responsible for revising the FSA's process for assigning risk ratings and compliance scores to food businesses and classifying them into risk categories (or 'segments'), explained that he had a problem about which he had come to ask my advice. He envisioned that his amended risk segmentation process would be partially automated, using algorithmic tools to assign businesses to appropriate risk categories without the need to subject them to a time consuming inspection. But a troubling question had occurred to him over the previous weekend: how would he know, once this algorithm had been developed, whether it was assessing risk and segmenting businesses correctly? How would he be able to tell whether his automated system was assigning the right businesses to the right risk categories? These were unfamiliar questions, with which Daniel had not had to grapple before. Did I have any suggestions about how the FSA might evaluate whether his new algorithm was 'working well'?

If I was a little confused about exactly what Daniel meant when he talked about an 'algorithm' in this context, I found his decision to come to me for help in evaluating its performance downright perplexing. I am still a newcomer to the problematics and politics of algorithmic governance, although this is rarely a handicap in my current work on food regulation. The regulators with whom I work – the FSA and the Local Authority Environmental Health Officers (EHOs)

[1] A pseudonym to protect this individual's anonymity.

and Trading Standards Officers (TSOs) who are primarily charged with enforcing food regulation in the UK – appear to interact with food businesses, and to evaluate and regulate the risks which they might pose, largely through conventional inspection practices. Officers carrying out inspections determine 'intervention rating' scores which express a business's level of compliance with food law and the magnitude of risk which it might pose to public health by assessing whether standardised lists of criteria have been met (for instance by taking temperature measurements, observing the hygiene practices of workers, and reviewing documentation). As such, the risk ratings and compliance scores which decide the frequency of future inspections are simultaneously both determined by the formalised logics of standardised inspection protocols and produced through the situational application of an inspecting officer's professional judgement and embodied skills.

The role of the latter often leads food businesses to complain that inspectors' decisions lack transparency and consistency, a charge partly responsible for provoking the changes to intervention rating and risk segmentation processes which Daniel is tasked with overseeing. The FSA has recently committed to developing a single centralised register of all food businesses operating in England, Wales and Northern Ireland. It is envisioned that this database (which at present remains an aspiration rather than a reality) will augment the FSA's existing records of inspection results, enforcement outcomes and results of tests on food samples with data sourced from other UK government business registers. This expanded dataset will then be used to classify (or 'segment') food businesses automatically into risk categories, helping regulatory bodies to distinguish 'high risk' businesses requiring immediate inspection from 'low risk' ones which may safely be inspected less frequently or even exempted from programmes of regular inspection altogether. This initiative thus aims to transform a regulatory system based on inspections conducted at regular intervals into a data-enabled system of what Karen Yeung (2016) might term 'pre-emptive enforcement' capable of using algorithmic systems to identify, and proactively investigate and intervene into, potentially risky businesses in order to prevent violations of food law.

Such proposals to use algorithms to assess risk, allocate regulatory resources and target intervention and enforcement action have attracted sustained critical scrutiny in recent years (for instance Janssen and Kuk 2016; Zarsky 2016). Some commentators fear that evacuating human judgement from risk profiling and the triggering of enforcement action might disperse responsibility across sprawling networks of designers, programmers, data providers and algorithm users – preventing victims of erroneous or disproportionate algorithmic judgements from locating a responsible party who can be held to account. Indeed, the technical complexity of algorithmic systems and their frequent seclusion from public scrutiny may render it difficult or impossible to establish the grounds on which a decision was made, and thus to determine whether it was correct or not. This opacity, combined with the speed with which algorithmic systems act and

the difficulty of reasoning with one, may make algorithmic decisions difficult to contest or appeal – a particularly worrying prospect given the potential for algorithmic systems to reproduce the biases of their designers or training datasets with potentially discriminatory results. It is surely important, in light of these concerns, to ask on what basis algorithms come to be authorised by regulatory actors to judge who is likely to comply with the law and who is likely to violate it. Through what tests and trials do algorithms become trusted, to put it in Daniel's terms, to make the right assessment?

It quickly became clear that Daniel had come to my desk already equipped with a preferred answer to this question. He proposed to test his risk segmentation algorithm for food businesses in a trial group of Local Authority areas over a period of perhaps one year. During this trial officers employed by participating Local Authorities would carry out a desktop risk assessment on each food business within their jurisdiction which fell due for inspection, using the same information and risk scoring rules as the algorithmic system. In this way it would be possible to compare the algorithm's assessment of the risk of non-compliance posed by each food business, and of whether or not each premises required inspection, against the professional judgements offered by Local Authority officers. Daniel's argument was that the FSA would know that the algorithm was 'working well,' even in the absence of comparable inspection records against which to test it (due to the introduction of his new risk segmentation process), if both algorithm and officers tended to identify the same premises as being high risk and in need of inspection. The more similar were the algorithm's conclusions to the judgements of the officers, the better it would be considered to be performing.

Daniel's suggested approach to testing and evaluation suggests something potentially interesting about the conditions under which algorithmic calculations might become accepted as valid judgements, and even as an acceptable basis on which to undertake intervention and enforcement action, within at least some regulatory institutions. For Daniel is proposing to evaluate the accuracy of his hypothetical algorithm's inferences about the riskiness and compliance status of food businesses through placing them in a very particular relationship with the professional judgements of human Local Authority officers. In his proposed trial, these assessors' expectations are to form the standard of 'sound judgement' against which the accuracy and efficacy of algorithmic decision making processes will be measured. Only if an algorithm can be shown to form roughly the same expectations about the probable riskiness and compliance status of food businesses as would a human desktop assessor can it be said to be 'working well' or trusted to 'make the right decision.' In other words, the use of algorithmic processes to target regulatory intervention can only be authorised once they have been shown to resemble the professional judgements of a trusted or authorised elite of human decision makers.

Evaluating the efficacy of algorithms in terms of the precision with which they mimic expert human judgement does not seem to be unusual in contemporary algorithm development practices. A few weeks before my conversation with Daniel, I had spoken to a supermarket food safety executive who described testing the effectiveness of his company's in-house risk mapping software through a similar process of validating its predictions about store risk scores against the 'gut feelings' of his staff. Meanwhile, Neyland and Möllers (2017) describe how developers of automated video surveillance systems evaluate their software's performance in detecting suspicious objects by comparing its propensity to issue security alerts against that of experienced CCTV operators watching the same footage. This suggests that even as the advent of algorithmic systems displaces human professionals from the making and execution of key regulatory decisions, it may in many cases also be establishing new relationships between human professional judgement and algorithmic decision processes. Perhaps paradoxically, these relationships sometimes appear to cast algorithmic judgement as an approximation of, and even as subordinate to, that of the very human professionals who algorithmic systems are intended to supplant. In Daniel's proposed trial, for instance, if an algorithm's predictions about which businesses present an elevated risk of non-compliance differ substantially from those of authorised human professionals then it will be judged to be 'performing poorly'. Only if the algorithm replicates the judgements of those professionals extremely closely will it stand a chance of being authorised to segment businesses into risk categories and to determine inspection frequencies in place of a human assessor.

While the testing and evaluation of algorithms may be distinctly contemporary practices, the epistemic hierarchies implicit in these trials also evoke some curious historical parallels. In some respects, Daniel's search for a mathematics which might replicate the operations of professional discrimination and judgement embodied in the decisions of Local Authority officers echoes the aspirations of eighteenth-century classical probability. This field of mathematics sought to convert the reasoning processes of esteemed subjects such as successful businessmen and distinguished magistrates into a 'calculus of good sense' which might be employed to guide the uninitiated towards wise courses of action (Daston 1988). Believing that the judgements of these 'reasonable men' expressed the best available estimates of the likelihood that, for instance, a given class of witness would testify truthfully in court or a business venture would be profitable, classical probabilists sought to derive mathematical descriptions of their deliberations. As such, the ultimate test of the classical probabilist's calculus was its degree of accuracy in describing and predicting the conduct of these 'reasonable men'. As in Daniel's proposed evaluation procedure for his risk segmentation algorithm, if a mathematical theory failed to predict the decisions of the 'reasonable men' then the theory – and not their judgment – was to be reassessed and reformulated.

Might these historical resonances provide some tentative insights into the possible implications of the apparent re-emergence of the aspiration to formalise the 'good sense' of an elite of authoritative human subjects into a mathematical calculus within the project of algorithmic regulation? As Lorraine Daston's (1988) magisterial history of eighteenth-century classical probability makes clear, its intellectual programme was riven by a host of epistemological, methodological and even political controversies – some of which may also return to haunt contemporary practices of algorithmic design, testing and authorisation. Who should be included in, or excluded from, the charmed circle of distinguished subjects whose reasoning is to form the standard against which the performance of algorithms and the accuracy of their decisions must be measured? How might this approach to evaluating algorithms and authorising their judgements accommodate disagreement among its referent group of reasonable human subjects about the correct decision or course of action under certain circumstances? How closely must an algorithm's judgements resemble those of this referent group in order for it to be performing 'well enough' that it may be authorised for use in real-world regulatory and enforcement practice? Indeed, how is this proximity to or distance from human judgement to be measured? To these historically familiar dilemmas might be added the questions of equity and fairness articulated by burgeoning social, legal and philosophical studies of algorithms. Might not a testing and evaluation process which defines a successful algorithm as one whose judgements mimic those of an elite group of authorised professionals produce algorithms which replicate this group's biases, much as other studies have suggested that algorithmic decision making may reproduce the prejudices of designers and programmers? Such an outcome might present cause for serious concern if it were to perpetuate existing inequities in official decision making through 'black boxing' these biases within opaque, distributed and secretive computational infrastructures whose operations may be difficult to understand, question or contest.

Although these concerns are familiar enough from other studies of algorithmic decision making, I would like to suggest tentatively that they manifest themselves in a specific form in this case due to the very particular relationship that Daniel's suggested approach to evaluation establishes between the judgements of authorised professionals and those of algorithms. Taking the judgements of a professional elite to constitute the standard against which the performance of algorithms must be measured transforms many issues of both equity and accuracy in algorithmic decision making into questions about the composition, consistency and impartiality of the group of human subjects in whose image algorithmic systems are being made. As such, the practices and protocols through which new algorithms are tested and their performance is evaluated are likely to be bound up intimately with the reconfigurations of authority, power and expertise which attend the introduction of algorithmic decision making processes into new arenas of regulatory practice. Indeed, the politics and ethics of algorithmic regulation may be negotiated as much through the shifting relations and processes of reference and validation into which evaluation processes draw

human and algorithmic judgement as through the formalisation of mathematical rules, the authorisation of data collection or the replacement of human assessors with automated systems. After all, decisions about how, and against which standards and measures, the performance of algorithmic systems should be evaluated are likely to play a crucial role in determining which algorithms are eventually authorised to remake regulatory processes. It may, then, be crucially important to investigate what might be at stake in the choice between different methods of determining whether an algorithm is making the 'right' decisions.

References

Daston, L. (1988) *Classical probability in the Enlightenment*, Princeton NJ: Princeton University Press.

Janssen, M. & Kuk, G. (2016) 'The challenges and limits of big data algorithms in technocratic governance', *Government Information Quarterly* 33(3): 371–77.

Neyland, D. and Möllers, N. (2017) 'Algorithmic IF … THEN rules and the conditions and consequences of power', *Information, Communication & Society* 20(1): 45–62.

Yeung, K. (2016) 'Algorithmic regulation and intelligent enforcement', in M. Lodge (ed.), *Regulation scholarship in crisis?,* CARR Discussion Paper no. 84, pp. 50–62. <http://www.lse.ac.uk/accounting/CARR/pdf/DPs/CARR_DP84-Martin-Lodge.pdf>

Zarsky, T. (2016) 'The trouble with algorithmic decisions: an analytic road map to examine efficiency and fairness in automated and opaque decision making', *Science, Technology, & Human Values* 41(1): 118–32.

Jeremy Brice, Centre for Analysis of Risk and Regulation, London School of Economics and Political Science.

EU data protection law falls short as desirable model for algorithmic regulation

Lee A. Bygrave

When canvassing various models for algorithmic regulation, EU data protection law is often posited as a rare example of such regulation operating in a legislative 'top-down' form and as a possible model to emulate. The main reason inheres in the fact that EU data protection law provides a qualified right for a person not to be subject to fully automated decisions based on profiling and supplements this with a right to knowledge of the logic involved in such decisions.

The right for a person not to be subject to fully automated decisions based on profiling currently inheres in Article 15 of the Data Protection Directive (95/46/EC). It is the first pan-European legislative norm aimed at regulating purely machine-based decisions in a data protection context. However, Article 15 has been, in practice, a second class right in the EU data protection framework. It has not been the subject of litigation before the Court of Justice of the EU (CJEU) or any national courts, bar those of Germany. Neither has it figured prominently in enforcement actions by national data protection authorities, nor in assessments of the adequacy of third countries' data protection regimes with a view to regulating the flow of personal data to these countries. This is not to say that the right has been simply symbolic, but the occasions in which it has been invoked appear to have been few and far between. While it has inspired a proposal for a similar right to be incorporated in a modernised version of the Council of Europe's 1981 Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data,[1] and also inspired various provisions of the 1997 Code of Practice on Protection of Workers' Data drafted by the International Labour Office (ILO),[2] it has not been widely replicated in the legal regimes of non-European countries.

Several features of Article 15 undermine its traction. Firstly, it only applies if four cumulative conditions are met: (i) a decision must be made; (ii) the decision must have legal or otherwise significant effects on the person whom the decision targets; (iii) the decision must be based solely on automated data processing; and (iv) the data processed must be intended to evaluate certain personal aspects of the person targeted by the decision. This creates multiple hurdles for the application of Article 15. Secondly, a considerable degree of ambiguity inheres in

[1] See Article 8(1) of the Draft modernised Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data [ETS 108], drawn up by the Council of Europe's Ad hoc Committee on Data Protection (version of September 2016). See too the Council of Europe's Guidelines on the Protection of Individuals with Regard to the Processing of Personal Data in the World of Big Data (adopted 23 January 2017; T-PD(2017)01), especially principles 7.1, 7.3 and 7.4.

[2] See particularly principles 5.5, 5.6, 6.10 and 6.11.

these conditions and this ambiguity is exacerbated by lack of authoritative guidance on how they are to be understood. Thirdly, even if all of the conditions for its exercise are met, the right is subject to fairly broad and nebulous derogations.

The General Data Protection Regulation (2016/679) will replace Directive 95/46/EC (DPD) in May 2018. Article 22 of the Regulation maintains the essence of the right provided by DPD Article 15, albeit in a somewhat different form.[3] Article 22 reads as follows:

> Article 22 Automated individual decision-making, including profiling
> 1. The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.
>
> 2. Paragraph 1 shall not apply if the decision: (a) is necessary for entering into, or performance of, a contract between the data subject and a data controller; (b) is authorised by Union or Member State law to which the controller is subject and which also lays down suitable measures to safeguard the data subject's rights and freedoms and legitimate interests; or (c) is based on the data subject's explicit consent.
>
> 3. In the cases referred to in points (a) and (c) of paragraph 2, the data controller shall implement suitable measures to safeguard the data subject's rights and freedoms and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision.
>
> 4. Decisions referred to in paragraph 2 shall not be based on special categories of personal data referred to in Article 9(1), unless point (a) or (g) of Article 9(2) apply and suitable measures to safeguard the data subject's rights and freedoms and legitimate interests are in place.

An important question is whether Article 22 will increase the power of EU data protection law over the generation and application of algorithms. My hunch is that it will not, for the following reasons. Firstly, Article 22, like its predecessor, still involves meeting multiple criteria. At the same time, Article 22 is afflicted to a greater degree than its predecessor by clumsy syntax that muddies its

[3] A similar right has also been inserted in Article 11 of the Directive on data protection and law enforcement (Directive (EU) 2016/680). Traces of the right (or, more accurately, an associated duty) are further found in Article 6 of the Directive on Processing of Passenger Name Record Data (Directive (EU) 2016/681).

interpretation. This clumsiness extends even to the very nature of the right provided by Article 22(1), which masquerades as a right of a data subject but is, in my opinion, really laying down a qualified prohibition on a particular type of decisional process.

Further, the derogations to the right in Article 22 have been, in one sense, broadened relative to the derogations under DPD Article 15. This means, for example, that automated decisions with the possibility of discriminatory effects, such as weblining, which are potentially hit by DPD Article 15, might be permitted under the new exception for consent. This exception is likely to lower the de facto level of protection for individuals, particularly in light of the relative strength of most individuals vis-à-vis banks, insurance companies, online service providers and many other businesses. However, the Regulation tightens the assessment of what is a *freely given* consent and what automated decisions are *necessary* for the purpose of entering into or performance of a contract (see Article 7(4) and recital 43). The traction of this tightening will rest on how strictly the necessity criterion is interpreted.

The level of protection under Article 22 will also depend on what safeguards the data controller is obliged to put in place under Article 22(3). In this respect, there is uncertainty as to whether Article 22(3) provides a right of ex post explanation of automated decisions – uncertainty that also afflicts interpretation of Arts. 14(2)(g) and 15(1)(h). However, there are aspects of Article 22(3) which offer a higher level of protection than under the DPD. According to Article 22(3), the data subject will *always* have the right to demand manual re-examination of a fully automated decision. This might take away the incentive for companies to acquire either a contract or consent from the data subject, as neither move will necessarily work as an exception to the right/prohibition in Article 22(1). Yet, if persons only rarely make use of their rights under Article 22(3), the overall effect of Article 22 on the automation of business might well end up being negligible – just as it has been with DPD Article 15.

The level of protection under Article 22 will further depend on the specifics of member state legislation, especially in light of the derogations provided in Article 22(2)(b) and, less directly, Article 22(4). At the same time, these derogations open up the possibility for significant divergence between national regulatory frameworks for automated decision making, thus undermining the harmonisation aims of the Regulation.

Finally, the traction of Article 22 will likely be weakened by practical difficulties in implementing its requirements, particularly in respect of decisional systems that are extremely complex and opaque, including for the controller(s). Such systems are increasingly the rule rather than the exception.

Lee A. Bygrave, Department of Private Law, University of Oslo.

Making sense of the European data protection law tradition

Karen Yeung*

Introduction

Although the UK Government Office for Science recently observed that identifying the right form of governance for artificial intelligence and for the use of digital data more widely is 'not self-evident' (Government Office for Science 2016), there is no doubt that contemporary data protection law will play a significant role. When academics and policymakers highlight the need for 'algorithmic accountability' their concern is not just with the software algorithms themselves, but the larger socio-technical systems in which those algorithms are embedded, including the data upon which those systems rely. Without data to sustain them, even the most sophisticated software algorithms are but hollow shells. Yet the so-called 'data deluge' precipitated by the digital transformation of economically advanced societies, and the dynamic and highly complex ecology within which personal data (i.e. data pertaining to identifiable individuals) now flows, has placed contemporary data protection law under unprecedented strain. Since its inception in the early 1970s, the contemporary western European data protection tradition has rested on the so-called 'data protection principles (see Bygrave 2014: ch. 5). Recent reform of EU Data Protection Directive provided European lawmakers with an opportunity to revisit the foundations of European data protection law, given that these core principles were formulated long before the internet had been invented. Yet the opportunity for radical reform was passed over, so that, although the EU General Data Protection Regulation (GDPR) introduces several regulatory innovations, the data protection principles continue to lie at its foundations. Hence Article 5 of the GDPR requires that personal data must be:

(a) Processed fairly, lawfully and in a transparent manner in relation to the data subject ('lawfulness, fairness and transparency')

(b) Collected for specified, explicit and legitimate purposes and not further processed for other purposes incompatible with those purposes ('purpose limitation');

(c) Adequate, relevant and limited to what is necessary in relation to the purpose for which data is processed ('data minimisation');

(d) Accurate and, where necessary, kept up to date ('accuracy');

* I am indebted to Lee Bygrave and Michael Veale for comments on earlier drafts.

(e) Kept in a form that permits identification of data subjects for no longer than is necessary for the purposes for which the persona data is processed ('storage limitation')

(f) Processed in a way that ensures appropriate security of the personal data including protection against unauthorised or unlawful processing and against accidental loss, destruction or damage, using appropriate technical or organisational measures ('integrity and confidentiality')

Yet many critics claim that the basic data protection principles, and the contemporary approach to European data protection law which they have spawned, is no longer fit for purpose. In order to evaluate this claim, we need to identify more clearly *what purposes* data protection law is designed to serve, and *how* European data protection law seeks to achieve those purposes. This paper aims to find answers to these questions, beginning first with the 'how' question by drawing insights from regulatory governance studies, before turning to the historical origins of modern European data protection in order to identify more clearly its point and purpose.[1] This will place us in a better position to evaluate whether or not European data protection law is 'fit for purpose' in an age of ubiquitous computing, a technological environment which differs radically from that which prevailed when European data protection laws first emerged.

How should we characterise the European approach to data protection?

Based on his thorough and thoughtful survey of data protection laws (which he refers to as 'data privacy law') throughout Europe, Lee Bygrave (2014: 2–4) draws attention to several characteristic features: it is largely statutory as these statutes often take the form of 'framework laws' that tend to set down diffusely formulated, general rules for processing, and thus allowing for the subsequent development of more detailed rules as the need arises; they usually establish independent bodies to oversee their implementation (referred to as 'data protection authorities' or 'privacy commissioners'), and these agencies often play a lead role in laying down how data privacy law is understood and applied, even in contexts where their views are only advisory. Although this is a very helpful starting point, for the regulatory governance scholar, it leaves many questions unanswered. Accordingly, the following discussion proceeds on the basis that the EU data protection tradition can be understood as a regulatory governance regime (or 'risk regulation regime') which regulates the collection and processing of personal data (Gellert 2015), enabling us to draw upon insights from regulatory governance studies to provide a deeper, more extensive account of the core strategies adopted in contemporary EU data protection law.

[1] I am indebted to Lee Bygrave, who, during the TELOS-CARR workshop on Algorithmic Regulation, questioned the validity of my claim that European data protection law adopted a 'rights-based' approach. Our ensuing discussion prompted me to write this paper.

A multi-instrument approach: command & control, the conferral of private rights and design-based techniques

There are a wide range of instruments and techniques that policymakers can utilise in seeking to influence social behaviour in pursuit of their policy goals. The basic framework of the EU data protection regime rests primarily upon a classical or 'command & control' strategy, by specifying a set of core legal standards (rooted in the basic data protection principles) which must be complied with by those who wish to collect and process personal data. Failure to comply with these legal requirements renders the resulting data handling activities unlawful (DPD Articles 7–8) exposing data controllers to enforcement action by national data protection authorities which may result in the imposition of significant civil penalties.[2] The new GDPR also introduces further prohibitions concerning the processing of children's data (Article 8 GDPR), and introduces new legal obligations requiring controllers to notify data protection authorities and data subjects of 'personal data breaches' (Articles 31-32 GDPR).

In addition to these basic prohibitions, the EU data protection regime also confers a set of rights on data subjects which data controllers are legally obliged to respect (and which data subjects may seek to enforce via the courts), including requirements that data controllers provide data subjects with basic information about the scope of data processing operations (DPD Articles 10-11), a series of access rights enabling the data subject to obtain knowledge of the logic involved in any automated processing of data concerning that individual (DPD Article 12(a), and a qualified right to object to certain types of fully automated decision-making processes (DPD Article 15(1). The new GDPR will also introduce new data subject rights, including a right to data portability (Article 18 GDPR). One further significant innovation in the techniques introduced by the GDPR are new requirements of 'data protection by design and by default' which, in essence, impose legal obligations on data controllers to 'hard wire' data protection norms into information systems development (GDPR Article 23). From a theoretical perspective, these provisions are very significant in acknowledging that the protection of digital data may be secured through 'design-based' regulatory techniques (Yeung 2008, 2015) by seeking to 'design in' normative standards into the artefacts, infrastructure and environment in which the regulated activities take place. But whether or not these measures lead in practice to a significant improvement in the level and comprehensiveness of protection remains to be seen.

In summary, EU data protection regime relies on several techniques and instruments, so that its overall approach is something of a cocktail, relying on a combination of conventional command & control techniques, typically applied by a public enforcement authority, but supplemented by a private rights regime and

[2] Under the GDPR, the penalties have increased to a maximum of €1 million or 2 per cent of an enterprise's annual turnover (per Article 79(6) GDPR).

recently bolstered by the introduction of 'design-based' strategies of protection (Morgan and Yeung 2007: ch. 2).

A process-based approach that eschews 'principles-based regulation'

Despite the multiplicity of regulatory strategies incorporated into the EU data protection regime, it is the legal prohibition on the collection and handling of personal data except in accordance with the data protection principles that provides the regime's central anchoring point. Yet to understand how these core principles are expected to function, we need to consider how they have been formulated, drawing on insights from regulatory governance scholars who have highlighted the relative merits of different approaches.

(a) Process vs outcome-based standards

Regulatory governance scholars often champion the benefits of performance or 'outcome-based' standards, which have become a well established feature of environmental regulation, in contrast to process-based standards (Lodge and Wegrich 2012: 15). The advantages of outcome-based standards are claimed to lie in conferring considerable flexibility and freedom on regulated parties in determining how they can meet a specified policy objective most effectively, thus avoiding the bluntness of process-based standards which require regulated firms to comply with a set of specified processes. In addition, performance against outcome-based standards is typically assumed to be measurable (such as air quality) and might therefore be less prone to circumvention or avoidance strategies by regulated parties. Despite these claimed advantages, the EU data protection principles are formulated largely in process-based terms, seeking to restrict the *processes* by and through which personal data must be handled, rather than specifying a set of *substantive values or outcomes* which data controllers must comply with when dealing with personal data (Gellert 2015).[3]

(b) Data protection principles rather than 'principles-based regulation'

Another central preoccupation arising in the regulatory governance literature focuses on challenges associated with standard-setting, including problems encountered in drafting appropriate standards to govern the regulated activity. A contrast is often drawn between detailed rules, on the one hand, and broadly drafted 'principles' on the other (Ford 2010). In his

[3] That said, some provisions of the GDPR specify substantive outcomes which must be achieved, sometimes in a fairly specific manner, such as Article 25(2) which provides that 'The controller shall implement appropriate technical and organizational measures for ensuring that, by default, only personal data which are necessary for each specific purpose of the processing are processed. That obligation applies to the amount of personal data collected, the extent of their processing, the period of storage and their accessibility. In particular, such measures shall ensure that by default personal data are not made accessible without the individual's intervention to an unspecified number of persons.' In addition, national legislative provisions derogating from the core norms of the GDP are only permissible to the extent that they provide for a substantive outcome as specified in Article 23(1).

penetrating analysis of the 'optimal precision of legal rules, Colin Diver observes that the success of a rule in effecting its purpose depends largely on several qualities: their transparency (whether the words have a well-defined and universally accepted meaning within the relevant community), their accessibility (applicable to concrete situations without excessive difficulty or effort), and their congruence with the underlying policy objective (whether the substantive content of the message communicated by the words produces the desired behaviour). Diver (1983) argues that rule drafters are inevitably confronted by the need to make trade-offs between these values. For example, detailed, narrowly specified rules may be highly transparent and accessible but may fail to achieve the desired behavioural outcome, and although broadly drafted principles may be highly congruent with the rule-maker's core policy objective, they often suffer from a lack of transparency and accessibility (Diver 1983).

The European data protection principles are drafted in fairly broad and general terms, and thus more appropriately described as principles rather than detailed rules. This does not however, imply that the EU data protection regime can be appropriately understood as a regime of 'principles-based regulation' (PBR). PBR refers to a regulatory approach that rose to prominence prior to the global financial crisis in the early to mid-2000s. Although PBR is not a term of art, it has been used by regulatory governance scholars to describe real-world regulatory regimes, which have tended not only to rely upon broadly based standards in preference to detailed rules, but also those regimes in which the standards themselves have also been largely *outcome-based* (rather than process-based) in their orientation (Black et al. 2007; Ford 2010). Accordingly, to describe the EU data protection regime as one of PBR would be to mischaracterise its approach, despite the central reliance on the EU data protection principles, due to the *procedural* nature of their demands, rather than prescribing a set of substantive outcomes or goals which data controllers are expected to meet in collecting and handling personal data.

Enforcement by a public regulator supplemented by private enforcement

One significant feature of the European data protection tradition is its reliance on independent administrative agencies, conventionally referred to as data protection authorities. Their primary responsibility is to oversee the operation of data protection legislation, including the task of investigating violations and taking enforcement action (including the imposition of financial penalties) against those found to be in breach. In addition to their broad discretionary powers of enforcement, data protection agencies also tend to wield considerable interpretive discretion, due to the relative dearth of established case law in which the courts have had an opportunity to clarify the scope and content of the legislative provisions (Bygrave 2014), and many of whom may promulgate administrative guidelines setting out how the agency proposes to interpret the law and exercise its enforcement discretion.

However, as noted above, EU data protection confers a number of legally enforceable rights on data subjects, which they can enforce against data controllers alleged to have violated these rights by taking court action seeking (among other things) an order that the infringement be terminated and the payment of damages. Yet in reality, although a majority of EU individuals are aware that they have some legal rights concerning their personal data (Hallinan et al. 2012), few are likely to be sufficiently motivated to initiate and maintain legal action against data controllers, particularly given the massive asymmetry in power, resources and sophistication of the giant digital platforms that now dominate our daily digital encounters. At the same time, although there are several grounds upon which a data controller may rely to demonstrate that the processing of personal data is lawful (a requirement of the 'lawful, fairness and transparency' principle), the consent of the data subject provides one such ground. Accordingly, by consenting to the collection, processing and transfer of personal data, data subjects may authorise controllers to use that data in more extensive and far-reaching ways than would otherwise be legally permitted. Having said that, because the purpose-limitation principle is unwaivable by data subjects, all data controllers are legally prohibited from processing data for purposes that are incompatible with the original purpose for which the data was collected. This does not, however, prevent data controllers from specifying the purpose of data collection in extremely broad terms, provided that those purposes are 'legitimate' and therefore fall within the controller's natural ambit of activity (see Bygrave 2002: 339–40). It is arguably due to the conferral of a set of legal rights on data subjects, combined with their power to consent to the collection and processing of their personal data by the controller, that one might be inclined to describe the general approach of EU data protection as 'rights based' in orientation. To the extent that there are enforceable legal obligations that prevent data controllers from collecting and processing personal data in accordance as they wish, then this is undoubtedly true. But it is misleading to the extent that such a description suggests that individual data subjects occupy a central role in monitoring and enforcing compliance with the regime.

Ex post vs ex ante regulation?

While the preceding discussion highlights the procedural nature of the EU data protection principles, this does little to illuminate *how* those principles are intended to foster the regime's underlying policy objectives. Although I seek to identify what, precisely, those policy objectives are in the next section, it is worth pausing to consider how the EU data protection regime would be understood in terms of a long-standing distinction, drawn primarily by economists, between ex ante and ex post approaches to regulation. Ex ante regulatory approaches typically rely on a system of prior approval, a technique which Anthony Ogus (1994: 214) observes can be traced back to the medieval guilds, that had effective monopolies over trades and crafts and to the issue of 'patents of monopoly' under royal prerogative, enabling certain individuals or groups to carry out trades or activities otherwise prohibited. In their modern guise,

regimes of this kind typically take the form of some kind of licensing regime, so that the regulatory activity is legally prohibited unless it is carried out by a valid licence holder, who has demonstrated to the relevant licensing authority that he or she has the requisite competence and capacities. Such approaches can be contrasted with ex post approaches, which typically entail the legal promulgation of certain minimum standards that the specified activity must meet, so that anyone who wishes to engage in that activity need not obtain prior permission, and may lawfully engage in it provided that the activity is undertaken in ways that meet the legislatively specified standards.

Because advance authorisation is generally not required before data controllers can lawfully collect and process personal data, the contemporary EU data protection regime can be understood as resting on an ex post rather than an ex ante strategy of control. Having said that, there is an ongoing debate within data protection circles about the relative merits of a so-called 'accountability' approach to data protection, touted as an alternative to conventional reliance on the data protection principles. In particular, Cate, Cullen and Mayer-Schönberger (2014), authors of the OECD 2013 revised guidelines, argue that because notice and consent is no longer an effective mechanism to protect the informational privacy of data subjects, it would be preferable to shift responsibility from the shoulders of individuals, who are currently required to weigh up their own interests in protecting privacy and accessing digital resources, in favour of an approach that reduces (or even eliminates) the purpose limitation and use limitation principles, thereby enabling largely unrestricted collection of personal data, but placing more onerous legal responsibilities on data controllers when seeking to process that data, requiring them to undertake a more focused evaluation of the risks and benefits associated with particular uses to ensure that harm to individuals is minimised (Cate et al. 2014). Despite criticism of this approach, the GDPR places greater emphasis on the concept of accountability than its predecessor, the EU Data Protection Directive.

Although this paper makes no attempt to evaluate the arguments put forward in support of this so-called accountability approach, its is a useful foil, highlighting the largely *preventive* approach reflected in the current structure and operation of the fair information processing principles. Hildebrandt (2015: 194) likens these principles to Odysseus's strategy of tying himself and his crew to the mast to prevent them responding to Sirens' call, thereby 'enabling them to resist the overweaning temptation to gather more and more data and use it for more and more intrusive purposes and applications that will ultimately lead to downfall and destruction'. Accordingly, although EU data protection does not institute an ex ante licensing regime, the general approach it takes towards the regulated activity (that of collecting and processing personal data) is largely a preventative one, aimed at averting the unlimited collection and re-purposing of personal data in order to reduce the dangers that might arise if there were no restrictions on the collection and use of personal data.

What are the substantive goals of European data protection law?

The preceding discussion draws attention to the largely preventive, process-oriented approach reflected in the European data protection tradition. Armed with this understanding of *how* contemporary European data protection laws are intended to function, we can now turn our attention to their overarching purpose(s). At a pragmatic level, the EU data protection regime (and basic data protection principles upon which it rests) was animated by a concern to strike an appropriate balance between the use of data and the protection of the interests of those to whom the data relate, in order to establish a set of agreed principles for handling personal data so that disparity in national standards would not impede the free flow of personal data across national borders and impede digital innovation. But what exactly are the interests of data subjects that require protection? In this respect, identifying in more precise terms the point and purpose of EU data protection is surprisingly elusive. As Bygrave (2014: 117) observes:

> data privacy law has long been afflicted by absence of clarity over its aims and conceptual foundations' and that this obscurity is reflected in the absence (in some privacy statutes) of the objects clauses formally specifying the interests that the legislation is intended to serve.

Nor can the objectives of European data protection law be adequately expressed solely in terms of the protection of informational privacy, although there is no doubt that the protection of informational privacy constitutes one important values which it seeks to protect.

In light of this lack of clarity, it is helpful to consider the historical origins of contemporary data protection law to identify the public anxieties that motivated their initial formulation. In providing a functional explanation for European data protection laws, Herbert Burkert (1981) articulates the 'problems' associated with data protection in terms of conflicts over the distribution of informational power wrought by new information communication technologies, which radically expanded the volume and speed of information that these technologies could handle, while also vastly expanding the number of individuals affected, and rendering temporal and spatial distance almost irrelevant in the acquisition and distribution of informational power. Drawing on Burkert's observations, Gellert (2015) interprets the history of European data protection law as fundamentally rooted in seeking to regulate the 'deployment of ICTs into society and, in particular, the data processing operations which they allow for' which 'put our freedoms at risk'. This resonates with Bygrave's observations that, in addition to privacy, there are a 'range of other interests which … form part of the rationale and agenda of data privacy law' including 'personal autonomy, integrity and dignity' and which he sums up as largely concerned with 'achieving individual goals of self-realisation' (Bygrave 2014: 119, citing Westin 1967). As Hildebrandt (2015: 191) puts it:

> The reason we need data protection is that putting together data has huge implications for a number of rights and freedoms. This is notably so when behavioural data are correlated.

One might wonder then, why the legal protection of the fundamental rights and freedoms that are accorded special status within the western European constitutional tradition was not regarded as capable of providing sufficient protection? While the answer to this question is highly speculative, it may lie partly in the way in which it is the *cumulative* and *aggregative* effects of Information and Communication Technology over time and space that threaten a number of fundamental rights and individual freedoms. Even before the arrival of the internet, the nature of digital data, coupled with the computational power and processes that enabled large bureaucratic institutions to collect personal data and compile and integrate it with other digital data sets to build up detailed informational profiles on individual citizens through automated processes that were many orders of magnitude cheaper, faster, and scalable than the processes for surveilling and profiling individuals in a pre-digital age, were recognised in Europe as a potentially serious threat to the individual rights and freedoms that are essential in thriving liberal democratic orders. Since the emergence of modern computing and associated ICTs, the rapid growth of their power and sophistication have delivered extraordinary benefits, many of which have become readily available to the wider population following the 'democratisation' wrought by personal computing and the widespread availability and take-up of smart connected devices. Indeed, contemporary life without the efficiency and convenience of the networked digital economy has become almost unthinkable. Yet these undeniable benefits can nevertheless serve to obscure the ways in which these technologies threaten to erode the social foundations upon which democratic freedom is rooted. Seen in this light, contemporary data protection law can be understood as analogous to environmental regulation, in seeking to protect to the democratic 'commons': rather than oriented towards protecting the natural, physical environment, it is oriented towards protecting the moral, democratic and cultural environment, by seeking to safeguard the collective social and cultural foundations which liberal democratic orders pre-suppose, and without which individual dignity, autonomy and self-development would not be possible. So understood, the need to establish a general regime of protection specifically concerned with limiting the collection and processing of personal data becomes more apparent.

The difficulty is, however, that the ways in which the collection and processing of personal data may threaten the democratic commons and the freedom, autonomy and dignity of individuals is not intuitively obvious, either to data controllers or to data subjects, particularly given the process-driven, preventive orientation of contemporary European data protection laws. Although the strong commitment to data protection in Europe relative to many other advanced industrialised nations can be attributed to relatively recent first-hand experience of totalitarian oppression, the memory of this experience may be fading with the passage of

time, particularly in light of the efficiency and convenience associated with digital tools that the collection and processing of personal data makes possible (Bygrave 2010). Although the EU constitutional framework, as amended by the Treaty of Lisbon, now recognises protection of personal data as a distinct, self-standing fundamental right (per Article 8 CFREU), thereby bolstering its normative significance, there is no instinctive or obvious connection between this right and the fundamental interests it protects. In contrast, other fundamental rights of a procedural nature, including the various procedural rights associated with the right to due process, including the right to a fair and public hearing, the right to a fair and impartial tribunal, and the right of an accused person to know the charges against her, have much more intuitive appeal, in that it is easy for ordinary citizens to recognise the vital interests which these procedural rights are concerned to protect. In contrast, the core interests and values which the fundamental right to data protection seeks to protect, and hence the normative and moral force with which we associate fundamental rights, are not so readily and instinctively evoked. Yet the right to data protection is justifiably accorded special status in view of its role in safeguarding the social foundations which make democratic society possible, and in which all individuals are treated with dignity and respect. As such, the complaint of Bert-Jaap Koops (2014) that European data protection law has failed to win the 'hearts and minds' of data controllers, who fail to recognise the substantive logic and rationale that underpin many data protection rules, can be understood. But it is far from self-evident that a regulatory regime that is oriented around the substantive values and interests upon which the existing data protection regime is rooted would fare any better. Such an approach might be more readily comprehensible to both data controllers and data subjects in terms of its underlying justification, couched perhaps in terms of the 'harms' associated with the collection and processing of personal data, might have considerable appeal. The GDPR can be understood as taking steps in this direction, to the extent that it places explicit reliance on the role of so-called 'data protection impact assessments'. But, for the time being at least, relying on these instruments to provide workable, legitimate and effective instruments to secure the protection of personal data that will offer clear and accessible guidance to data controllers concerning the content and limits of permissible data handling while nurturing public trust seems naively optimistic.

Conclusion

In order to understand the role and potential of European data protection law in securing the accountability of algorithmic systems which rely on the processing of personal data, it is necessary to identify how these laws are intended to operate, and what they are intended to do. This analysis has highlighted several features of the European legal regime. In particular, I have suggested that its preventive, process-oriented nature, which seeks to restrict the way in which personal data is collected and processed in order to prevent excessive 'data power' accumulating in the hands of data controllers, makes it difficult for both data subjects and data controllers to intuitively recognise the underlying substantive rights, interests and values which the regime is ultimately aimed at

protecting. Although placing greater emphasis on data processing 'harm' represents movement towards a more substantive approach, whether or not this will serve to win public hearts and minds by enabling them to grasp the importance of the need to protect the social foundations of democratic orders which these regimes seek to safeguard is far from guaranteed.

References

Burkert, H. (1981) 'Institutions of data protection – an attempt at a functional explanation of European national data protection laws.' *Computer Law Journal* 3: 166–88.

Bygrave, L.A. (2002) *Data protection law: approaching its rationale, logic and limits*. London: Kluwer Law International.

Bygrave, L.A. (2010) 'Privacy and data protection in an international perspective', *Scandinavian Studies in Law,* pp. 166–200.

Bygrave, L.A. (2014) *Data privacy law.* Oxford: Oxford University Press.

Black, J., Hopper, M. and Band, C. (2007) 'Making a success of principles-based regulation', *Law and Financial Markets Review* 1: 191–206.

Cate, F., Cullen, P. and Mayer-Schönberger, V. (2014). *Data protection principles for the 21st century: revising the 1980 OECD guidelines*. Redwood VA: Microsoft. (https://www.oii.ox.ac.uk/archive/downloads/publications/Data_Protection_Principles_for_the_21st_Century.pdf)

Diver, C. (1983) 'The optimal precision of legal rules', *Yale Law Journal* 93: 65–109.

Ford, C. (2010) 'Principles-based securities regulation in the wake of the global financial crisis', *McGill Law Journal* 55: 257–307.

Gellert, R. (2015) 'Data protection: a risk regulation? Between the risk management of everything and the precautionary alternative', *International Data Privacy Law* 5: 3–19.

Government Office for Science (2016) *Artificial Intelligence: opportunities and implications for the future of decision-making*. London: Government Office for Science.

Hallinan, D., Friedewald, M. and McCarthy, P. (2012) 'Citizens' perceptions of data protection and privacy in Europe', *Computer Law and Security Review* 28(3): 263–72.

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

Koops, B-J. (2014) 'The trouble with European Data Protection Law', *International Data Privacy Law* 4: 250–61.

Lodge, M. and Wegrich, K. (2012) *Managing regulation*. London: Palgrave Macmillan.

Ogus, A.I. (1994) *Regulation: legal form and economic theory*. Oxford: Clarendon.

Morgan, B. and Yeung, K. (2007) *An introduction to law and regulation*, Cambridge: Cambridge University Press.

Westin, A.F. (1967) *Privacy and freedom*, New York: Atheneum Press.

Yeung, K. (2008) 'Towards an understanding of regulation by design', in R. Brownsword and K. Yeung (eds), *Regulating technologies: legal futures, regulatory frames and technological fixes,* Portland OR: Hart Publishing.
Yeung, K. (2015) 'Design for regulation', in J. van den Hoven, I. van de Poel and P. E. Vermaas (eds), *Handbook of ethics, values and technological design*, Dordecht: Springer.

Karen Yeung, Centre for Technology, Ethics, Law & Society (TELOS), King's College London.

carr
centre for analysis
of risk and regulation

Centre for Analysis of Risk and Regulation
The London School of Economics and Political Science
Houghton Street
London WC2A 2AE
Email: risk@lse.ac.uk
lse.ac.uk/CARR

TELOS
http://www.kcl.ac.uk/law/research/centres/telos/index.aspx