THE RULE OF LAW AND AUTOMATION OF GOVERNMENT DECISION-MAKING

MONIKA ZALNIERIUTE, LYRIA BENNETT MOSES AND GEORGE WILLIAMS

(2019) 82(3) *Modern Law Review*
[2019] *UNSWLRS* 14

UNSW Law
UNSW Sydney NSW 2052 Australia

Pre-edited version, (2019) 82(3) *Modern Law Review*

The Rule of Law and Automation of Government Decision-Making

Monika Zalnieriute,[*] Lyria Bennett Moses[**] and George Williams[***]

Governments around the world are deploying automation tools in making decisions that affect rights and entitlements. The interests affected are very broad, ranging from time spent in detention to the receipt of social security benefits. This article focusses on the impact on rule of law values of automation using: (1) pre-programmed rules (for example, expert systems); and (2) predictive inferencing whereby rules are derived from historic data (such by applying supervised machine learning). The article examines the use of these systems across a range of nations. It explores the tension between the rule of law and rapid technological change and concludes with observations on how the automation of government decision-making can both enhance and detract from rule of law values.

INTRODUCTION

Automation promises to improve a wide range of processes. The introduction of controlled procedures and systems in place of human labour can enhance efficiency as well as certainty and consistency. Given this, it is unsurprising that automation is being embraced by the private sector in fields including pharmaceuticals, retail, banking and transport. Automation also promises benefits to government. It has the potential to make governments – and even whole democratic systems – more accurate, more efficient and more fair. As a result, several nations have become enthusiastic adopters of automation in fields such as welfare allocation and the criminal justice system. While not a recent development, automated systems that support or replace human decision-making in government are increasingly being used.

The rapid deployment of automation is attracting conflicting narratives. On the one hand, the transformative potential of technologies such as machine learning has been lauded for its economic benefits. On the other, it has become customary to acknowledge the risks that these pose to rights such as privacy[1] and equality.[2] The question of how automation

[*] Postdoctoral Research Fellow, Allens Hub for Technology, Law and Innovation, Faculty of Law, UNSW Sydney.
[**] Director, Allens Hub for Technology, Law and Innovation, Faculty of Law, UNSW Sydney.
[***] Dean, Anthony Mason Professor and Scientia Professor, Faculty of Law, UNSW Sydney; Barrister, New South Wales Bar. The authors thank Gabrielle Appleby and the anonymous referees for their comments on an earlier draft, and Adam Yu and Leah Grolman for their research assistance.

[1] For automation, data protection and privacy, see, eg, A. Roig, 'Safeguards for the Right Not to be Subject to a Decision Based Solely on Automated Processing (Article 22 GDPR)' (2017) 8 *European Journal of Law and Technology* 1; S. Wachter, B. Mittelstadt and L. Floridi, 'Why a Right to Explanation of Automated Decision-Making does not Exist in the General Data Protection Regulation' (2017) 7 *International Data Privacy Law* 76; S. Wachter, B. Mittelstadt and C. Russell, 'Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR' (2017) 31 *Harvard Journal of Law & Technology* 841; I. Mendoza and L. A. Bygrave, 'The Right Not to Be Subject to Automated Decisions Based on Profiling' in T. Synodinou et al (eds), *EU Internet Law: Regulation and Enforcement* (Cham: Springer: 2017); G. Malgieri and G. Comandé, 'Why a Right to Legibility of Automated Decision-Making Exists in the General Data Protection Regulation'

interacts with foundational legal concepts and norms is also attracting attention among theorists working at the intersection of legal theory, technology and philosophy.[3] These scholars examine possibilities such as the potential of automation and artificial intelligence to displace traditional legal concerns with prediction,[4] and indeed to challenge the normative structure underlying our understanding of law.[5] Others have interrogated the relationship between legal values and data-driven regulation.[6] Another area of focus is 'artificial legal intelligence' and its potential for improving access to justice and to provide benefits for historically marginalised populations.[7] These and other questions are typically examined in particular legal or factual contexts, such as in regard to administrative law or law enforcement.[8]

(2017) 7 *International Data Privacy Law* 243; B. Goodman and S. Flaxman, 'European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation"' (2017) 38 *AI Magazine* 50. See also UN Office of the High Commissioner for Human Rights (OHCHR), *A Human Rights-Based Approach to Data: Leaving No One Behind in the 2030 Development Agenda* (2016); United Nations Development Group, *Big Data for Achievement of the 2030 Agenda: Data Privacy, Ethics and Protection – Guidance Note* (2017) at https://undg.org/document/data-privacy-ethics-and-protection-guidance-note-on-big-data-for-achievement-of-the-2030-agenda/ (last accessed 27 November 2018).

[2] For automation and equality, see, eg, S. Barocas and A. D. Selbst, 'Big Data's Disparate Impact' (2016) 104 *California Law Review* 671; M. B. Zafar et al, 'Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment' (International World Wide Web Conferences Steering Committee, 2017) *Proceedings of the 26th International Conference on World Wide Web* at https://dx.doi.org/10.1145/3038912.3052660 (last accessed 10 September 2018); A. Chouldechova, 'Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments' (2017) 5 *Big Data* 153; S. Goel et al, 'Combatting Police Discrimination in the Age of Big Data' (2017) 20 *New Criminal Law Review* 181. See also 'The Toronto Declaration: Protecting the rights to equality and non-discrimination in machine learning systems' 16 May 2018 at https://www.accessnow.org/the-toronto-declaration-protecting-the-rights-to-equality-and-non-discrimination-in-machine-learning-systems/ (last accessed 27 November 2018)

[3] See, eg, recent special issue 'Artificial Intelligence, Technology, and the Law' (2018) 68 supp 1 *University of Toronto Law Journal* 1, focused on legal theory, automation and technology beyond government decision-making. See also K. Yeung, 'Algorithmic Regulation: A Critical Interrogation' (2017) *Regulation & Governance* at https://doi.org/10.1111/rego.12158 (last accessed 10 September 2018); A. Rouvroy and B. Stiegler, 'The Digital Regime of Truth: From the Algorithmic Governmentality to a New Rule of Law' A. Nony and B. Dillet (tr), 2016, 3 *La Deleuziana* 6 at http://www.ladeleuziana.org/wp-content/uploads/2016/12/Rouvroy-Stiegler_eng.pdf (last accessed 10 September 2018); E. Benvenisti, 'EJIL Foreword – Upholding Democracy Amid the Challenges of New Technology: What Role for the Law of Global Governance?' (2018) 29 *European Journal of International Law* 9; M. Hildebrandt and B. Koops, 'The Challenges of Ambient Law and Legal Protection in the Profiling Era' (2010) 73 MLR 428.

[4] F. Pasquale and G. Cashwell, 'Prediction, Persuasion, and the Jurisprudence of Behaviourism' (2018) 68 supp 1 *University of Toronto Law Journal* 63.

[5] M. Hildebrandt, 'Law as Computation in the Era of Artificial Legal Intelligence: Speaking Law to the Power of Statistics' (2018) 68 supp 1 *University of Toronto Law Journal* 12; B. Sheppard, 'Warming Up to Inscrutability: How Technology Could Challenge Our Concept of Law' (2018) 68 supp 1 *University of Toronto Law Journal* 36, 37; M. Hildebrandt, *Smart Technologies and the End(s) of Law: Novel Entanglements of Law and Technology* (Cheltenham: Edward Elgar, 2015).

[6] M. Hildebrandt, 'Profiling and the Rule of Law' (2008) 1 *Identity in the Information Society* 55; F. Pasquale, 'Toward a Fourth Law of Robotics: Preserving Attribution, Responsibility, and Explainability in an Algorithmic Society' (2017) 78 *Ohio State Law Journal* 1243; D. K. Citron and F. Pasquale, 'The Scored Society: Due Process for Automated Predictions' (2014) 89 *Washington Law Review* 1.

[7] P. Gowder, 'Transformative Legal Technology and the Rule of Law' (2018) 68 supp 1 *University of Toronto Law Journal* 82.

[8] In the context of administrative decision-making, see, eg, M. Oswald, 'Algorithm-Assisted Decision-Making in the Public Sector: Framing the Issues using Administrative Law Rules Governing Discretionary Power' (2018) 376 *Philosophical Transactions of the Royal Society A* 20170359 at https://doi.org/10.1098/rsta.2017.0359 (last accessed 10 September 2018); C. Coglianese and D. Lehr, 'Regulating by Robot: Administrative Decision Making in the Machine-Learning Era' (2017) 105 *Georgetown Law Journal* 1147; D. Hogan-Doran, 'Computer Says "No": Automation, Algorithms and Artificial Intelligence

This article adopts a broader perspective in assessing the benefits and challenges to the rule of law posed by automation of government decision-making.[9] The goal is not to provide an exhaustive analysis, but to critically investigate how principles of the rule of law are affected by the increasing use of two kinds of automation: human-authored pre-programmed rules (such as expert systems) and tools that derive rules from historic data to make inferences or predictions (often using machine learning). Our focus in doing so is on three core rule of law concepts that have the widest acceptance across political and national systems: transparency and accountability; predictability and consistency; and equality before the law.

These rule of law values are applied to four case studies: automated debt-collection in Australia, data-driven risk assessment by judges in the United States, social credit scoring in China, and automated welfare in Sweden. The case studies have been selected to provide a diverse range of viewpoints from which to assess the benefits and risks to the rule of law posed by the use of automated decision-making by governments around the world. We do not provide a detailed consideration of jurisdiction-specific constitutional, administrative and statutory requirements constraining decision-making in these nations.[10] Our aim instead is to analyse developments at the conceptual level of how they impact upon the rule of law, rather than seeking to develop a detailed prescription for the design or implementation of such systems.

We conclude that the alignment of automated government decision-making with rule of law values hinges on the appropriateness of design choices. The most significant factor is whether the automated system uses explicit rules written by humans (generally to align with legal requirements for the relevant decision) or rules derived empirically from historic data to make inferences relevant to decisions or to predict (and thus mimic) decisions. The latter raise greater issues for transparency and accountability, particularly as newer techniques are often more complex and therefore less susceptible to human explanation. Further, such systems are less likely to be consistent with the law and more likely to fall foul of the principle of equality before the law. In practice, however, systems of both types can fail to live up to rule of law ideals. The solution lies in ensuring that system design reflects rule of law values which are appropriate to the kind of decision being supported or made.

---

in Government Decision-Making' (2017) 13 *Judicial Review* 345. In the context of national security and law enforcement, see, eg, L. Bennett Moses and L. de Koker, 'Open Secrets: Balancing Operational Secrecy and Transparency in the Collection and Use of Data for National Security and Law Enforcement Agencies' (2017) 41 *Melbourne University Law Review* 530; Hildebrandt, n 6 above; T. Z. Zarsky, 'Transparent Predictions' [2013] *University of Illinois Law Review* 1503.

[8] See, eg, M. Hildebrandt and S. Gutwirth (eds), *Profiling the European Citizen: Cross-Disciplinary Perspectives* (Dordrecht: Springer, 2008); Hildebrandt, n 6 above; D. Lyon, 'Surveillance, Snowden, and Big Data: Capacities, Consequences, Critique' (2014) 1 *Big Data & Society* 1; P. De Hert and S. Gutwirth, 'Privacy, Data Protection and Law Enforcement. Opacity of the Individual and Transparency of Power' in E. Claes, A. Duff and S. Gutwirth (eds), *Privacy and the Criminal Law* (Antwerpen & Oxford: Intersentia, 2006); A. D. Selbst, 'Disparate Impact in Big Data Policing' (2017) 52 *Georgia Law Review* 109.

[9] A few short commentaries exist calling for more attention to be paid to the governmental context: see, eg, S. J. Mikhaylov, M. Esteve, and A. Campion, 'Artificial Intelligence for the Public Sector: Opportunities and Challenges of Cross-sector Collaboration' (2018) 376 *Philosophical Transactions of the Royal Society A* 20170357 at https://doi.org/10.1098/rsta.2017.0357 (last accessed 10 September 2018); R. Kennedy: 'Algorithms and the Rule of Law' (2017) 17 *Legal Information Management* 170; M. Perry, 'iDecide: Administrative Decision-Making in The Digital World' (2017) 91 *Australian Law Journal* 29.

[10] For example, in the United States, this would include due process protections in the Administrative Procedure Act, Pub L 79-404, 60 Stat 237, 5 USC §§ 551-559.

RULE OF LAW

The rule of law is a political work in progress, at the heart of which lies a widely held conviction that society should be governed by law. The prominence of the rule of law is such that diverse societies and seemingly irreconcilable political regimes, ranging from the European Union to Russia, China, Zimbabwe and Iran, have endorsed the concept. Some of these societies reject democracy and human rights, others oppose capitalism and globalisation, and some defy liberalism and are openly anti-Western,[11] but they all embrace an ideal of the rule of law.

Acceptance of the rule of law across so many nations and political systems is possible because the concept lacks an accepted definition. It is ubiquitous, yet elusive. As an 'essentially contested concept',[12] different societies can endorse the rule of law while disagreeing about what it entails. As Tamanaha notes:

> Some believe that the rule of law includes protection of fundamental rights. Some believe that democracy is part of the rule of law. Some believe that the rule of law is purely formal in nature, requiring only that laws be set out in advance in general, clear terms, and be applied equally to all.[13]

At the highest level of abstraction, Tamanaha recognises that 'the rule of law is analogous to the notion of "good," in the sense that everyone is for it, but having contrasting convictions about what it is.'[14]

Some scholars have separated understandings of the rule of law into *formal* and *substantive* conceptions. The former focuses on sources and forms of legality, while the latter also includes stipulations about the content of the law.[15] The idea that the rule of law embodies both procedural and substantive elements is widely accepted.[16] For example, Lord Bingham argued that the core principle of the rule of law is 'that all persons and authorities within the state, whether public or private, should be bound by and entitled to the benefit of laws publicly and prospectively promulgated and publicly administered in the courts'.[17] He further articulated eight core principles, including accessibility and predictability, application of law, equality of law, protection of fundamental rights, availability of civil disputes proceedings, limits on power exercised by public officials, fairness of adjudicative procedures provided by the state, and state compliance with its obligations under international law.[18] Lord Bingham's articulation of the rule of law is a further attempt to expound a concept that, by its nature, defies universal definition.

It is not our goal to provide yet another account of the rule of law.[19] Instead, we focus narrowly on aspects of the rule of law that have general acceptance, notably that it requires governance in which the law must be predictable, stable, accessible and everyone must be

---

[11] B. Z. Tamanaha, *On the Rule of Law: History, Politics, Theory* (Cambridge: Cambridge University Press, 2004) 2.

[12] J. Waldron, 'The Concept and the Rule of Law' (2008) 43 *Georgia Law Review* 1, 52. See also S. Sedley, *Lions under the Throne: Essays on the History of English Public Law* (Cambridge: Cambridge University Press, 2015). On essentially contested concepts more generally, see W. B. Gallie, 'Essentially Contested Concepts' in M. Black (ed), *The Importance of Language* (Ithaca, NY: Cornell University Press, 1962) 121.

[13] Tamanaha, n 11 above, 3.

[14] *ibid*, 3.

[15] P. P. Craig, 'Formal and Substantive Conceptions of the Rule of Law: An Analytical Framework' [1997] PL 467.

[16] See, *ibid*, 467.

[17] Lord Bingham, 'The Rule of Law' (2007) 66 CLJ 67, 69.

[18] *ibid*.

[19] Modern accounts include Lord Bingham, n 17 above; Tamanaha, n 11 above; P. Gowder, *The Rule of Law in the Real World* (Cambridge: Cambridge University Press, 2016).

equal before the law.[20] In applying these principles, our focus is primarily upon the formal and procedural aspects of the rule of law, rather than its capacity to encompass a broader set of human rights, including to free speech and privacy. Hence, we limit our analysis to the following core components: transparency and accountability; predictability and consistency; equality before the law.

Transparency and accountability

One of the best-known aspects of the rule of law is that governments must be transparent and accountable in respect of the rules and decisions they make. Transparency requires publicity about the operation of the state and that individuals can access legal rules and administrative decisions.[21] This is important so that individuals can understand the reasons for decisions affecting them and learn how future decisions might affect them. In democratic systems, some awareness as to the principles underlying the operation of the law (albeit not necessarily the specific details of decisions affecting others) is also useful for people seeking to understand and hence evaluate the performance of government. Accountability further requires that government be subject to the law and answerable for its actions (for example, that executive action can be overturned where it transgresses the law).[22] Transparency and accountability are related because the transparency of a decision-making process or system is necessary (but not sufficient) for making that process or system accountable.[23] This includes accountability as to compliance with other rule of law principles, such as equality before the law.

Predictability and consistency

Another widely accepted aspect of the rule of law is that the law should be predictable and consistent.[24] Many regard this as indispensable for individual freedom and a fundamental part of 'what people mean by the Rule of Law'.[25] Predictability and consistency of law is often thought to have dual purpose. It enhances certainty and efficiency so that individuals may

[20] Report of the International Congress of Jurists, 'The Rule of Law in a Free Society' (New Delhi: International Commission of Jurists, 1959) at [1].

[21] See, Gowder, n 19 above.

[22] R. Mulgan, *Holding Power to Account: Accountability in Modern Democracies* (New York. NY: Palgrave Macmillan, 2003); A. Schedler, 'Conceptualizing Accountability' in A. Schedler, L. Diamond and M.F. Plattner (eds), *The Self-Restraining State: Power and Accountability in New Democracies* (Boulder, CO: Lynne Rienner, 1999) 17.

[23] Bennett Moses and de Koker, n 8 above, 534–537.

[24] L. L. Fuller, *The Morality of Law* (New Haven, CT: Yale University Press, 1964); see also the lists in J. Finnis, *Natural Law and Natural Rights* (Oxford: OUP, 2nd ed, 2011) 270–271; J. Rawls, *A Theory of Justice* (Oxford: OUP, 1999) 208–210; J. Raz, *The Authority of Law: Essays on Law and Morality* (Oxford: OUP, 1979) 214–218.

[25] M. Schwarzschild, 'Keeping it Private' (2007) 44 *San Diego Law Review* 677, 686. For example, in his well-known book on the subject, Tom Bingham indicated that one of the most important things people needed from the law that governed them was predictability in the conduct of their lives and businesses. He quoted Lord Mansfield to the effect that: '[i]n all mercantile transactions the great object should be certainty: … it is of more consequence that a rule should be certain, than whether the rule is established one way rather than the other.': *Vallejo* v *Wheeler* (1774) 1 Cowp 143, 153 cited in T. Bingham, *The Rule of Law* (London: Allen Lane, 2010) 38. Similarly, Paul Gowder has recently argued that one of the main requirements for a political state under the Rule of Law is *regularity:* those who use state coercion must actually be bound by reasonably specific legal rules in that use: P. Gowder, 'Transformative Legal Technology and the Rule of Law' (2018) 68 *University of Toronto Law Journal* 82, 89 (summarising the main boundaries of the rule of law in his work, Gowder, n 19 above). See also F. A. von Hayek, *The Constitution of Liberty* (Chicago, Ill: University of Chicago Press, 1960).

manage their private lives and affairs effectively.[26] It also has a moral significance in that like cases ought to be treated equally, an issue explored separately below. In common law systems, predictability and consistency of law are furthered by judicial adherence to precedent.[27]

Equality before the law

Equality before the law stipulates that all human beings must be subject to and treated equally by the law without inappropriate reference to their status or other circumstances. This implies due process, including that all individuals are subject to the same rules of justice.[28] The question as to whether due process is an aspect of equality before the law or a separate principle is open to debate;[29] we analyse it here as an aspect of equality before the law. In broad terms, equality before the law might guarantee that no individual or group be privileged or discriminated against due to their racial or ethnic background, sex, national origin, religious belief, sexual orientation or other irrelevant personal characteristics.[30] In this form, equality before the law might give rise to a range of substantive rights, though the scope and content of these remain contested.[31] In this article, we apply a narrow conception of equality before the law, that is, that people, irrespective of their status, must have equal access to rights in the law and that, in accessing these rights, 'like cases be treated alike'.[32] We adopt this approach without making any claim that this is exhaustive of the principle. This narrow conception of equality before the law has the broadest possible application for a range of legal and political systems.

AUTOMATION OF DECISION-MAKING

Automation in government decision-making is not a new phenomenon, nor is it linked to a single technology. If a human government decision-maker were to automatically decide

[26] See Bingham, n 17 above; see also W. Eskridge and P. Frickey (eds), *Hart and Sacks's The Legal Process: Basic Problems in the Making and Application of Law* (Westbury, NY: Foundation Press, 1994).
[27] J. Waldron, 'Stare Decisis and the Rule of Law: A Layered Approach' (2012) 111 *Michigan Law Review* 1; D. A. Farber, 'The Rule of Law and the Law of Precedents' (2005) 90 *Minnesota Law Review* 1173. S. A. Lindquist and F. C. Cross, 'Stability, Predictability and the Rule of Law: *Stare Decisis* as Reciprocity Norm' University of Texas Law School, Conference Paper, 2010 at https://law.utexas.edu/conferences/measuring/The Papers/Rule of Law Conference.crosslindquist.pdf (last accessed 15 August 2018).
[28] Egalitarian moral value is attached to this principle by all theorists who argue that the principle is part of the conception of the rule of law: see, eg, A. V. Dicey, *Introduction to the Study of the Law of the Constitution* (Indianapolis: Liberty, 8th ed, 1982) 114–115; Waldron, n 27 above; von Hayek, n 25 above, 85, 209. For a classical liberal work on equality before the law, see A. L. Hudson, 'Equality Before the Law' (1913) CXII *The Atlantic Monthly* 679.
[29] See, eg, J. Waldron, 'The Rule of Law and the Importance of Procedure' (2011) 50 *Nomos* 3.
[30] For broad, substantive accounts of rule of law, see R. Dworkin, *Law's Empire* (Cambridge, MA: Belknap, 1986); Gowder, n 19 above, chs 2–3.
[31] For examples of minimalist positions, see J. Rousseau, *The Social Contract* (1762, M. Cranston tr, London: Penguin, 2003), Fuller, n 24 above, ch 2; J. Raz, 'The Rule of Law and its Virtue' in J. Raz, *The Authority of Law: Essays on Law and Morality* (Oxford: Clarendon Press, 1979) 214–218; J. Finnis, *Natural Law and Natural Rights* (Oxford: OUP, 2nd ed, 2011) 270–271; C. R. Sunstein, *Legal Reasoning and Political Conflict* (New York, NY: OUP, 2018) 119–122; M. J. Radin, 'Reconsidering the Rule of Law' (1989) 69 *Boston University Law Review* 781.
[32] See Rawls, n 24 above, 237; H. L. A. Hart, 'Positivism and the Separation of Law and Morals' (1958) 71 *Harvard Law Review* 593, 623–624. The notable exception is Raz, for whom the rule of law does not include principle of equality before the law, see Raz, n 24 above.

every case in the same way, one might say that the decision-maker acted as an automaton. This would also be the case if the decision-maker applied a simple criterion, such as approving only those applications made before lunchtime. By contrast, the automation examined in this article relies on a technological tool or system. The extent to which it does so varies along a spectrum (of partial through to full automation) from decision support (computer helps humans make decisions) to human-in-the-loop (decisions are made with some human involvement) to the disappearance of humans from the decision-making process entirely.

To illustrate, it is useful to consider concrete examples moving from decision-making where automation plays a supporting role to decisions that are made entirely by machines. Starting from decision-support tools, a facial recognition tool used by a customs official at an airport might identify an applicant as being on a security watchlist and pull up the record of that person from a database. The official might then review information in the database, question the applicant, and decide whether to admit that person to the country. Further along the spectrum are systems that automatically determine some fact relevant to a decision, such as that an individual meets an age criterion, while leaving the remaining elements of decision-making to a government official. Still further, an automated system might provide information relevant to an evaluative rather than purely fact-based criterion, such as assessing whether an individual is likely to be dangerous or unlikely to comply with a payment plan. Automation might also recommend that the decision-maker decide a case a particular way, in which case the decision-maker may treat such a recommendation as more or less determinative of the outcome. Finally, a system may identify the relevant information, and then make a decision based upon that information without engaging a human decision-maker. This might occur, for example, in determining whether an applicant has met the criteria to receive a welfare benefit.

There are a variety of technologies that are being used, or are likely to be used, to automate government decision-making processes over the immediate term. In analysing the impact of automation on the rule of law, we divide these into two classic types, although we recognise that these can also be combined in a decision-making process. The first type of automation is a process that follows a series of pre-programmed rules written by humans. The second type of automation deploys rules that are inferred by the system from historic data. Before explaining how these might be combined, it is worth exploring each separately through the lens of examples – expert systems and supervised machine learning, respectively.

Expert systems are sometimes described as the first wave of artificial intelligence,[33] a general term used to describe situations where machines perform tasks that would ordinarily require human intelligence. They are an example of a pre-programmed logic where rules are coded into a system and applied to new examples to reach a conclusion. Typically, these rules are written by, or designed in consultation with, those who have sufficient knowledge of the domain in which the decision will operate;[34] for example, in the context of government-decision-making, those with knowledge of the relevant legislative provisions and decision criteria. Expert systems can be used to automate components of a decision-making process that rely on clear, fixed and finite criteria. If legislation provides that individuals who meet criteria A, B and C are eligible for a benefit, an expert system can operate so that only individuals meeting all three of those criteria (with inputs coming from responses to a questionnaire, from a government database and/or from some other source) receive the benefit.

[33] See, generally, A. Tyree, *Expert Systems in Law* (Sydney: Prentice Hall, 1989).
[34] D. A. Waterman and M. A. Peterson, *Models of Legal Decisionmaking: Research Design and Methods* (Santa Monica, CA: Rand, 1981) 13-14.

Because they are based on explicit rules, expert systems can give reasons for decisions, citing the material facts and rules on which a conclusion was reached. In the extremely simple example of an expert system that determines whether criteria A, B, and C are met, the output of the system might read 'Applicant X is not eligible for this benefit because criterion C is not met'. The reasons for the decision can be generated because the system is using the same logic as the rule itself, namely assessment of each applicant against criteria A, B and C. The system can also be rendered transparent to the public by writing the encoded rules as one or more statements, such as 'An applicant is eligible for the benefit if and only if A, B and C are met'.[35]

Expert systems have been used by governments both to augment and to replace human decision-makers. Since the 1980s, such systems have been designed for use in a variety of government contexts, such as child protection and calculation of welfare benefits.[36] Robo-debt and the Swedish welfare system, discussed below, are more modern examples of systems using a pre-programmed human-authored logic. Whether or not they are coded as traditional 'expert systems' (which typically separate the inference engine from the rules database), they mirror a similar approach. In particular, they operate on the basis of human-crafted logic, with identical inputs inevitably yielding the same output.

Quite different are systems that automate decision-making, not on the basis of explicit human-authored rules, but on the basis of rules learnt from patterns and correlations in historic data. Machine learning, which falls into the 'second wave' of artificial intelligence,[37] automates the construction of the rules that drive the system. Machine learning describes a variety of data-driven techniques that establish processes by which a system will 'learn' patterns and correlations so that it can generate predictions or reveal insights. The learning occurs iteratively as an algorithm attempts to improve performance against a specified goal.

Supervised machine learning requires data that has already been classified or labelled, for example as to whether in that circumstance an applicant is eligible or not eligible for a benefit. Because the data is pre-labelled (either in the context of historic decision-making or in the context of development of the system), it carries within it human biases and assumptions.[38] For example, crime data may reflect policing and judicial biases towards minority groups, while data on eligibility for benefits may reflect bureaucratic impulses to reduce spending. Those deploying supervised machine learning must also decide how they wish to evaluate performance (for example, false positives might be preferred to false negatives). The process typically begins by dividing the data (whatever its source) into a

---

[35] R. E. Susskind, *Expert Systems in Law: A Jurisprudential Inquiry* (Oxford: Clarendon Press, 1987) 114–115.

[36] eg, J. R. Schuerman et al, 'First Generation Expert Systems in Social Welfare' (1989) 4 *Computers in Human Services* 111; J. Sutcliffe, 'Welfare Benefits Adviser: A Local Government Expert System Application' (1989) 4 *Computer Law & Security Review* 22.

[37] J. Launchbury, 'A DARPA Perspective on Artificial Intelligence' DAPRAtv, YouTube, 2017 at https://www.youtube.com/watch?v=-O01G3tSYpU (last accessed 20 August 2018). The Defence Advanced Research Projects Agency (DARPA) has also named a third wave of artificial intelligence that has not been applied to government decision-making and so is not explored further in this paper.

[38] eg, L. Gitelman (ed), *'Raw Data' is an oxymoron* (Cambridge, MA: MIT Press, 2013); S. Barocas and A. D. Selbst, 'Big Data's Disparate Impact' (2016) 104 *California Law Review* 671; T. Calders and I. Žliobaitė, 'Why unbiased computational processes can lead to discriminative decision procedures' in B. Custers, T. Calders, B. Schermer, and T. Zarsky (eds), *Discrimination and privacy in the information society: Data mining and profiling in large databases*. (Heidelberg: Springer, 2013) 43–60; R. Kitchin, *The Data Revolution: Big Data, Open Data, Data Infrastructures & Their Consequences* (London: Sage Publications, 2014); B. E. Harcourt, *Against prediction: Profiling, policing and punishing in an actuarial age* (Chicago, Ill: University of Chicago Press, 2013); J. Lerman, 'Big Data and its Exclusions' 66 *Stanford Law Review Online* 55 at https://www.stanfordlawreview.org/online/privacy-and-big-data-big-data-and-its-exclusions/ (last accessed 27 November 2018).

training set and a testing set – the latter being reserved to evaluate the performance of the algorithm according to the relevant criteria. The rule that is learnt can then be applied to the testing data to evaluate the algorithm, after which further adjustments may be made.

There are different methods that might be used in 'training' the machine, offering various levels of comprehensibility among other features. A supervised machine learning process may learn a simple rule (for example, that eligibility hinges on the presence of factors and the absence of others) or it may learn a 'rule' that involves an extended series of steps for which there is no apparent logic. The kind of rule that is learnt will hinge on the model of machine learning that is deployed as well as the kinds of patterns existing in the data. Eventually, it is deployed on real world data during decision-making.

Supervised machine learning is an example of a broader range of methods that aim to draw inferences from data for the purposes of drawing an inference or making a prediction. Other techniques, including those associated with traditional statistics, can be used to achieve a similar end. For example, a regression analysis can be used to estimate relationships among variables, which can be used to write a rule for predicting a particular variable (such as the outcome of a decision). However, unlike standard statistical methods, machine learning is generally iterative (capable of continually 'learning' from new information) and capable of identifying more complex patterns in data.

The line between the two types of automation (pre-programmed and rules derived from historic data) is not always clear. Humans can write explicit rules that are based not on statutory criteria or legal doctrine, but rather on empirical findings gleaned from historic data (through statistics or machine learning). In this case, a rule is inferred from data at a particular point in time but then is pre-programmed into a system. A system that automatically follows the same rule, originally learnt through a machine learning process, has some characteristics of each of expert systems and machine learning. Like expert systems, it cannot operate outside its programmed parameters. If the rule becomes obsolete (for example, because statutory criteria or decision-making policies change), it will no longer be effective at predicting decisions that would be made by humans. The system will also share with machine learning the potential for complexity (discussed in relation to Transparency and Accountability, below). In particular, depending on the machine learning process employed, the rule generated (and used) may hard for humans to understand, explain or justify. This example demonstrates that our two types of automation are not strictly separate categories. Nevertheless, they are useful 'classic types' that help to distinguish different kinds of challenges that arise for the rule of law. Where relevant, we discuss the possibility of blending the two types of automation as a solution to particular rule of law challenges.

Despite automating the decision-making process to varying extents, none of the approaches to automation considered here remove humans from the process entirely. Humans decide which processes to automate and what techniques to deploy, as well as identify data or rules that will form the basis for inferences. For example, in the context of supervised machine learning, it is generally[39] humans who decide key matters such as what will be predicted and how this will be measured, what data is collected and whether and how errors are corrected.[40] At least at this stage of technological development, most of the automation comes after humans have designed and built the system. This means that the human aspect of these technologies can never be discounted.

[39] Potentially, if artificial intelligence becomes more sophisticated, machines will become involved in these processes. But for now, they remain under the control of humans.

[40] D. Lehr and P. Ohm, 'Playing with the Data: What Legal Scholars Should Learn about Machine Learning' (2017) 51 *University of California Davis Law Review* 653.

CASE STUDIES OF AUTOMATION IN DECISION-MAKING

As a reference point for our analysis in the following section, we describe below programs where governments are relying on automation in making decisions that affect individuals. These case studies represent a diverse selection of nations and technological approaches as well as different stages of implementation.

Robo-debt in Australia

Robo-debt is a nickname given by the media to a controversial program, announced by the Australian government in 2015, to calculate and collect debts owed because of welfare overpayment.[41] It replaced a system of manual review of individuals selected through risk management, where income and other information was gathered from the individuals, their bank records and employer records.

Under the robo-debt system, data on annual income held by the Australian Tax Office (ATO) was automatically cross-matched with income reported to the government welfare agency Centrelink. Because welfare entitlements were originally calculated on the Centrelink figure, a higher income declared to the ATO was taken to mean that the individual concerned had been overpaid and thus owed a debt to the government. The system thus combined data matching (possibly employing machine learning),[42] automated assessment through the application of human-authored formulae, and automated generation of letters to welfare recipients.

To understand how the system worked, it is important to know that income is reported to the ATO as an annual figure but to Centrelink as a fortnightly figure. The first step was to check the two annualised income figures against each other. Where the ATO annual income was greater than the Centrelink annualised income, individuals were sent a letter giving them an opportunity to confirm their annual income through an online portal. Those who accessed the online portal were given an opportunity to state their *fortnightly* income (with evidence), whereas those who did not access the portal were assumed to earn a fortnightly figure calculated as the annual ATO figure divided by the number of weeks in a year.[43] However, the letter sent to individuals did not explain that recording variation in income over the year was important to an accurate calculation of welfare entitlements.[44] The fortnightly income (entered into the online system or derived as above) was used to calculate what the welfare entitlement ought to have been and, where relevant, individuals were automatically sent a debt notice. Some letters were sent to individuals who did not in fact owe any money because variations in their income were not recorded and had an impact on their welfare entitlements. While the system has been modified over time, our comments are here directed to its original implementation. Several concerns have been raised in the use of this system. These include poorly worded correspondence, inaccuracy of the formula in a percentage of cases, issuing debt notices to those not owing money,[45] shifting the burden of proof,[46] and leaving individuals to the mercy of debt collectors.[47]

---

[41] The program was introduced as part of a 2015–16 Budget measure, 'Strengthening the Integrity of Welfare Payments' and a December 2015 Mid-Year Economic Fiscal Outlook announcement.

[42] Such data matching is authorised by the Data Matching Program (Assistance and Tax) Act 1990 (Cth).

[43] Commonwealth Ombudsman, 'Centrelink's Automated Debt Raising and Recovery System: A Report about the Department of Human Services' Online Compliance Intervention System for Debt Raising and Recovery' (Investigation Report, 2017) 1, 4 at https://www.ombudsman.gov.au/__data/assets/pdf_file/0022/43528/Report-Centrelinks-automated-debt-raising-and-recovery-system-April-2017.pdf (last accessed 27 November 2018).

[44] *ibid*, 9.

[45] T. Carney, 'The New Digital Future for Welfare: Debts without Legal Proofs or Moral Authority?' UNSW Law Journal Forum, May 2018 at http://www.unswlawjournal.unsw.edu.au/wp-content/uploads/2018/03/006-Carney.pdf (last accessed 16 August 2018).

Data-driven risk assessment in US sentencing decisions

In some jurisdictions in the United States, judges use an automated decision-making process called COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) that draws on historic data to infer which convicted defendants pose the highest risk of re-offending, particularly where there is a risk of violence. The Conference of Chief Justices in the United States has come out in support of judges using such tools, including in the sentencing process.[48] Such use has also been endorsed by the Supreme Court of Wisconsin in *Wisconsin* v *Loomis*[49] (*Loomis*). That case held that partial reliance on a COMPAS score in sentencing (affecting the non-parole period of a sentence) did not violate the defendant's right to due process under the United States Constitution. The Court found that such reliance is permissible so long as the decision is not fully delegated to the output of the machine learning software – for example, a judge will still need to consider a defendant's arguments as to why other factors might impact the risk they pose.[50] On the other hand, there is no requirement that defence counsel be able to challenge the accuracy of the COMPAS tool or the algorithms upon which it is based, both of which remain a trade secret.[51]

Risk assessment tools such as COMPAS distinguish among individuals based on a variety of characteristics. The full extent of these are not known given the proprietary nature of the software. Concerns have been raised that race has an impact on assessments. For example, a ProPublica investigation found that African Americans are more likely than whites to be given a false positive score by COMPAS.[52] This is not necessarily because race is used as a variable in modelling relative dangerousness of the offender population; differential impact can result where race correlates with variables that are themselves correlated with risk classification. Differential outcomes can thus result where the data on which the system is trained is itself steeped in human biases.

While racial discrimination was not an issue in *Loomis*, gender discrimination was raised. Data on gender was included in the set on which the algorithm was trained, the reason being that rates of re-offending, particularly violent re-offending, differ statistically between men and women. The Supreme Court of Wisconsin held that this kind of differential treatment did not offend the defendant's due process right not to be sentenced based on his male sex. Its reason was that because men and women have different rates of recidivism, ignoring gender would 'provide less accurate results'.[53] This highlights a fundamental question about the logic employed in drawing inferences using rules derived from historic data – if the goal is to maximise predictive accuracy, does it matter from a rule of law perspective whether individuals are classified differently based on inherent characteristics?

---

[46] P. Hanks, 'Administrative Law and Welfare Rights: A 40-Year Story from *Green v Daniels* to "Robot Debt Recovery"' (2017) 89 *AIAL Forum* 1, 9–11.

[47] Note that this aspect of the program has been modified, see Commonwealth Ombudsman, n 3 above at [1.35], [1.48], [3.16].

[48] CCJ/COSCA Criminal Justice Committee, 'In Support of the Guiding Principles on Using Risk and Needs Assessment Information in the Sentencing Process' (Resolution 7, adopted 3 August 2011) at http://ccj.ncsc.org/~/media/Microsites/Files/CCJ/Resolutions/08032011-Support-Guiding-Principles-Using-Risk-Needs-Assessment-Information-Sentencing-Process.ashx (last accessed 15 August 2018).

[49] *State of Wisconsin* v *Loomis* 881 N.W.2d 749 (Wis. 2016). The United States Supreme Court denied certiorari on 26 June 2017.

[50] *ibid* at [56].

[51] *ibid* at [51].

[52] J. Angwin et al, 'Machine Bias' ProPublica, 23 May 2016 at https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (last accessed 16 August 2018).

[53] *Loomis* n 49 above at [77], [86].

Automated student welfare in Sweden

The Swedish National Board of Student Finance (CSN) has been singled out by the Swedish government as a pioneer in the use of automated decision-making by public agencies.[54] The CSN manages financial aid to students in Sweden for their living costs, which includes grants and various loans.[55] The core target group of the CSN generally has high knowledge of, and access to, information technologies. The CSN automated rule-based decision-making system is mandated by national legislation, and the role of professional officers is to guide customers through the e-service in accordance with an ethical code.[56] This ensures that the decisions are based on clear, public rules and a human confirms and takes responsibility for each decision.

The automated system is available both to potential applicants for student loans and grants (managed by a so-called 'out' unit), as well as those who are paying their loans back to the CSN (managed by the 'in' unit).[57] Numerous e-services provided by CSN are partially or fully automated. For example, an e-service that allows people to apply for a reduction in repayments is used to support decision-making process (partial automation), while all the decisions on loan re-payments based on income of the last two years are fully automated. The automated decision-making system combines data from CSN with publicly available information, including tax information (which is publicly available in Sweden).[58] Whenever an individual applies for a reduction, an officer enters any relevant information into the system manually before letting the automated system take over again, meaning that the system is partially automated. While it is the system that 'makes' decisions, the officers are obliged by law to take responsibility for them and to communicate the decisions to the customers by editing the default formulation and signing it.

Social credit system in China

A fourth case study of automation is the Social Credit System (*shehui xinyong tixi* – SCS) developed by central government in China and implemented by 43 'demonstration cities' and districts at a local level.[59] According to the government planning document that outlines the system,

> its inherent requirements are establishing the idea of a sincerity culture, and promoting honesty and traditional virtues, it uses encouragement for trustworthiness and constraints against untrustworthiness as incentive mechanisms, and its objective is raising the sincerity consciousness and credit levels of the entire society.[60]

---

[54] Näringsdepartementet, Statens Offentliga Utredningar, 'En digital agenda i människans tjänst [A digital agenda in the service of people]' Statens Offentliga Utredningar [Official Reports of the Swedish Government], Report no SOU 2014:13, 2014 at https://www.regeringen.se/rattsliga-dokument/statens-offentliga-utredningar/2014/03/sou-201413/ (last accessed 16 August 2018).

[55] See the website of the CSN at https://www.csn.se/languages/english.html (last accessed 6 November 2018).

[56] E. Wihlborg, H. Larsson, and K. Hedström. '"The Computer Says No!" A Case Study on Automated Decision-Making in Public Authorities' 2016 49th Hawaii International Conference on System Sciences.

[57] See the CSN website, n 55 above.

[58] Swedish Tax Agency, 'Taxes in Sweden: An English Summary of Tax Statistical Yearbook of Sweden' 2016 at https://www.skatteverket.se/download/18.361dc8c15312eff6fd1f7cd/1467206001885/taxes-in-sweden-skv104-utgava16.pdf (last accessed 10 September 2018).

[59] A linguistic note made by Rogier Creemers is useful in this context: 'the Mandarin term "credit" (*xinyong*) carries a wider meaning than its English-language counterpart. It not only includes notions of financial ability to service debt, but is cognate with terms for sincerity, honesty, and integrity.': see R. Creemers, 'China's Social Credit System: An Evolving Practice of Control' 2018 at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3175792 (last accessed 16 August 2018).

[60] R. Creemers (ed), 'Planning Outline for the Construction of a Social Credit System (2014–2020)' (Eng tr of State Council Notice of 14 June 2014) 25 April 2015 at

In accordance with such goals, the SCS provides rewards or punishments as feedback to individuals and companies, based not just on the lawfulness, but on the morality of their actions, covering economic, social and political conduct.[61]

From a technological perspective, the SCS resembles a straightforward, pre-programmed rule-based system, however each of 43 'model cities' implement the programme differently. For example, under the Rongcheng City model,[62] everyone is assigned a base score of 1,000 points on a credit management system, which connects four governmental departments. Subsequent points are then added or deducted on the system by (human) government officials for specific behaviour, such as, for example, late payment of fines or traffic penalties. There are in total 150 categories of positive conduct leading to additional points on the system, and 570 categories of negative behaviour leading to point deductions for individuals. The implications of the SCS cover a wide range of economic and social repercussions. For instance, those with low social credit rating scores may not be eligible for loans and certain jobs, or may be denied the opportunity to travel on planes or fast trains. In contrast, those with high scores enjoy benefits such as cheaper public transport, free gym facilities and priority for shorter waiting times in hospitals.

The SCS is still in its early stages and the Chinese government has been forming partnerships with private companies with sophisticated data analytics capacity. For example, the central government has been cooperating with Chinese tech giant Alibaba in a Sesame Credit system, which includes, among other things, an automated assessment of potential borrowers' social network contacts in calculating credit scores.[63] This means that those with low-score friends or connections will see a negative impact on their own scores because of an automated assessment.[64] Sesame Credit combines information from the Alibaba database with other personal information, such as individual browsing and transaction history online, tax information and traffic infringement history, to automatically the determine the trustworthiness of individuals.

BENEFITS AND CHALLENGES TO THE RULE OF LAW

Transparency and accountability

Automation offers many potential benefits in enhancing the transparency and accountability of governmental decision-making. Whereas a human may come up with justifications for a decision *ex post* that do not accurately represent why a decision was made,[65] a rules-based system can explain precisely how every variable was set and why each conclusion was reached. It can report back to an affected individual that the reason they were ineligible for a benefit was that they did not meet a criterion that is a requirement of a legislative or

---

https://chinacopyrightandmedia.wordpress.com/2014/06/14/planning-outline-for-the-construction-of-a-social-credit-system-2014-2020/ (last accessed 16 August 2018).

[61] For a detailed analysis of thinking and design process behind the SCS, see Creemers, *ibid*.

[62] 荣成：建信用体系 创"示范城市 [Rongcheng: The Making of a Demonstration City for the Social Credit System] 新华社 [Xinhua News Agency], 13 July 2017 at http://xinhua-rss.zhongguowangshi.com/13701/6003014383535113117/2049163.html (last accessed 10 September 2018).

[63] M. Hvistendahl, 'Inside China's Vast New Experment in Social Ranking' *Wired*, 14 December 2017 at https://www.wired.com/story/age-of-social-credit/ (last accessed 10 September 2018).

[64] R. Zhong and P. Mozur, 'Tech Giants Feel the Squeeze as Xi Jinping Tightens His Grip' *New York Times* (online), 2 May 2018 at https://www.nytimes.com/2018/05/02/technology/china-xi-jinping-technology-innovation.html (last accessed 10 September 2018).

[65] R. E. Nisbett and T. DeCamp Wilson, 'Telling More Than We Can Know: Verbal Reports on Mental Processes' (1977) 84 *Psychological Review* 231.

operational rule that is pre-programmed into the logic of the system. It is important to note here that such feedback is not *necessarily* provided for rules-based expert systems. The designer decides what the output of the system will be and whether it will include reasons for its conclusions or decisions. In the case of robo-debt, individuals were not provided with clear information as to how the debts were calculated in general, or in their individual case. The opposite is true for the Swedish system, where decisions are made based on clear, public rules and a human confirms and takes responsibility for each decision.

To understand the barriers to transparency, it is helpful to understand Burrell's three 'forms of opacity'.[66] The first form is intentional secrecy, which arises when techniques are treated as a trade or state secret, or when data used in the process contains personal information which cannot be released due to privacy or data protection laws. This form of opacity can apply to systems based on rule-based logic and systems that derive rules from data using techniques such as machine learning. In the case of the Chinese Social Credit system, only limited information is made public. For example, the details of the cooperation between the central government and the private sector in the Sesame Credit system are not clear. While it is known that the system will use machine learning and behavioural analytics in calculating credit scores,[67] individuals have no means of knowing what information from their social network contacts was used or its precise impact on their scores.[68]

A government agency may also outsource the building of or licence the use of an automated system and will then be bound by contractual terms that prevent further disclosure.[69] In the case of COMPAS, Northpointe Inc (now 'equivant'),[70] which built the tool, has not publicly disclosed its methods as it considers its algorithms trade secrets.[71] While the risk assessment questionnaire and thus the input variables have been leaked,[72] there is insufficient information available about methods and datasets used in training. The lack of transparency was the focus of one of the concurring judgments in *Loomis*.[73] Abrahamson J noted that 'this court's lack of understanding of COMPAS was a significant problem in the instant case' and that 'making a record, including a record explaining consideration of the evidence-based tools and the limitations and strengths thereof, is part of the long-standing, basic requirement that a circuit court explain its exercise of discretion at sentencing.'[74] Such transparency and analysis of the tool itself would also, in her opinion, provide 'the public with a transparent and comprehensible explanation for the sentencing court's decision'.[75]

While trade secret rights may legitimately be claimed by private corporations, and enforced against contracting parties who agree to confidentiality provisions, there are important questions from the perspective of the rule of law about whether secret systems can

[66] J. Burrell, 'How the Machine "Thinks": Understanding Opacity in Machine Learning Algorithms' (2016) 3 *Big Data & Society* 1.
[67] Hvistendahl, n 63 above.
[68] Zhong and Mozur, n 64 above.
[69] For a discussion of intellectual property rights limiting the transparency of algorithms, see G. Noto La Diega, 'Against the Dehumanisation of Decision-Making: Algorithmic Decisions at the Crossroads of Intellectual Property, Data Protection and Freedom of Information' (2018) 9 JIPITEC 3, 11–16. In the context of outsourcing, there are additional considerations (beyond non-transparency) that may have legal implications that are beyond the scope of this paper.
[70] 'Equivant' at http://www.equivant.com/ (last accessed 10 September 2018).
[71] This is noted in *Loomis* n 49 above at [144].
[72] See Angwin, n 52 above.
[73] *Loomis* n 49 above.
[74] *ibid* at [133], [141].
[75] *ibid* at [142].

be used in government decision-making in contexts that directly affect individuals. In at least some circumstances, rule of law considerations should favour open source software.

The second form of opacity identified by Burrell, again potentially relevant to both kinds of automation considered here, is technical illiteracy.[76] Here, the barrier to greater transparency is that even if information about a system is provided (such as a technique used in training a machine learning algorithm or the formal rules used in an expert system), most people will not be able to extract useful knowledge from this. A system may accordingly be transparent to a technical expert, while remaining opaque to the majority of the governed, including those affected by particular decisions. Of course, those without specialist knowledge can consult those with it, just as those affected by badly drafted laws may need to consult with lawyers in order to understand their obligations. However, in some contexts, particularly where the consequences of a decision are severe, the lack of access to expert advice in understanding and challenging a decision effectively reduces the extent to which the decision itself can be described as transparent and accountable in practice.

The third form of opacity that Burrell describes relates specifically to machine learning and stems from the difficulty of understanding the action of a complex learning technique working on large volumes of data, even equipped with the relevant expertise.[77] For example, the process through which a face is 'recognised' by an automated system may involve a complex combination of distal relationships, angles, colouring, shape and so forth, combined through a multi-layered neural network, each layer reflecting different combinations of multiple variables. Whereas the second form of opacity involved limitations of expertise, the third form of opacity recognises human limitations in truly understanding or explaining the operation of complex systems. Because humans reason differently to machines, they cannot always interpret the interactions among data and algorithms, even if suitably trained. This suggests that the transparency necessary for the rule of law may decrease over time as machine learning systems become more complex.

There are some possible and partial solutions to this challenge. Some researchers are working on 'explainable AI', also known as XAI, which can explain machine learning inferences in terms that can be understood by humans.[78] It is also possible to disclose key information about a machine learning system, such as the datasets that were used in training the system and the technique that was used. Machine learning systems can also be made transparent as to aspects of their operation. Evaluations and testing can be used to ensure that systems satisfy stated requirements, whether based on predictive accuracy or equal treatment of groups. In other words, the use of automation can be justified or explained by a decision-maker based on its empirically observed properties rather than on its inputs and methods. For example, if one has a question about an algorithmic process, such as whether it discriminates against a group, one can use tools that test for this without disclosing algorithmic methods or data sources more broadly.[79] This qualified transparency can at least ensure that outputs are accountable along particular dimensions (such as compliance with equality standards).

However, some machine learning techniques cannot be rendered transparent, either generally, in particular circumstances or to particular people. The three challenges identified

[76] Burrell, n 66 above, 4.
[77] *ibid* 5, 10.
[78] For example, there is an XAI program at the Defence Advanced Research Projects Agency in the US that aims to develop machine learning systems that 'will have the ability to explain their rationale, characterise their strengths and weaknesses, and convey an understanding of how they will behave in the future.': D. Gunning, 'Explainable Artificial Intelligence (XAI)' (Defense Advanced Research Projects Agency Project Information) at https://www.darpa.mil/program/explainable-artificial-intelligence (last accessed 16 August 2018).
[79] J. A. Kroll et al, 'Accountable Algorithms' (2017) 165 *University of Pennsylvania Law Review* 633.

by Burrell, taken together, mean that there will rarely be public transparency as to the full operation of a machine learning process, including understanding reasons for the decision, understanding limitations in the dataset used in training (including systemic biases in the raw or 'cleaned' data), and accessing the source code of the machine learning process. In some cases, it may be sufficient that particular information about an algorithm (its equal treatment of different groups, for example) is rendered transparent through evaluation and testing. However, there are circumstances where qualified or limited transparency may be insufficient from a rule of law perspective. The use of the COMPAS system in sentencing, which ultimately impacts on individual liberty, is an example of a situation where a high degree of transparency is needed to comply with rule of law values.

An alternative solution lies in the fact that decision-making systems only need to be transparent and accountable *as a whole*, which does not necessarily imply visibility of the entire operation of automated components of that system. For example, in the Swedish student welfare example and elsewhere,[80] a human remains accountable for the decision, even though the logic itself is first run through an automated system. Ultimately, the success of this strategy depends on its implementation. If the human can be called on to provide independent reasons for the decision, so that the automated system is essentially a first draft, then the decision-making system as a whole is as accountable and transparent as it would have been in the absence of decision-support software. If, however, the human can rely on the output of the system as all or part of their reason for the decision, then accountability for the decision remains flawed despite assurances. This goes back to the question of the degree of automation in the decision-making process and the influence of outputs over the ultimate decision. A decision-making system as a whole can be made transparent and accountable by marginalising automated components (at the cost of efficiency and other benefits) and ensuring human accountability in the traditional way or by rendering transparent and accountable those automated components.

As is evident from the above, the degree of transparency inherent in an automated system is a question of human design choices. The system designer can choose what information about the decision-making process to output. And the bureaucracy determines the role of the automated system within the broader context of decision-making. While some methods are more difficult to render transparent, it is the choice of the designer as to whether such methods are used at all in particular systems. There are constraints – as Burrell points out, machine learning tools are often opaque whether due to deliberate policy (of government or a private contractor), lack of expertise in the community, or complexity of the method selected. This means that there may be compromises needed between transparency and choice of software or tool. The best predictor may not be the most transparent or may be difficult to situate in a system of accountability.

Thus, where decisions are fully or partially automated, the transparency and accountability of outputs hinges on the accountability of those designing the system *for* the transparency and accountability of the decision-making system itself. Indeed, a similar point is true for all rule of law values. They are unlikely to be found in decision-making and decision-support systems by accident. Those designing systems should be required to design

[80] For example, the Home Affairs Department Secretary in Australia has stated that 'no robot or artificial intelligence system should ever take away someone's right privilege or entitlement in a way that can't ultimately be linked back to an accountable human decision-maker': D. Wroe, 'Top Official's "Golden Rule": In Border Protection, Computer Won't Ever Say No' *Sydney Morning Herald* 15 July 2018 at https://www.smh.com.au/politics/federal/top-official-s-golden-rule-in-border-protection-computer-won-t-ever-say-no-20180712-p4zr3i.html (last accessed 13 January 2019).

them in ways consistent with the rule of law (including the criteria analysed here) and be able to give an account of how this has been done.

Many of the humans involved in designing systems and setting relevant parameters are data scientists, computer scientists and engineers. Professionally, there has been a move to the development of standards, frameworks and guidelines to ensure that decision-making and decision-support systems are *ethical*.[81] This suggests another potential way forward for the rule of law, writing it into the language of technical specifications for decision-making and decision-support systems deployed by government. Designers could then be made accountable for meeting those standards, whether contractually, professionally or through regulation. The challenge of converting an essentially contested concept into technical specifications (in one or multiple versions) would not be an easy one, and we do not attempt it here.

However it is achieved, the need for greater transparency about automated decision-making software, its development, and the assumptions embedded, as well as the weighting of different variables by such systems, is one of the most frequently emphasised issues by both technical and legal experts.[82] It is also crucial from the perspective of the rule of law. Firstly, it could lead to greater understanding of these systems, the values underlying them and their operation, thus revealing what is now obscure. More transparency would also allow affected individuals to challenge such decision-making systems, because information about the variables, inputs and outputs would be available.[83] For example, Citron and Pasquale have developed a concept of 'technological due process', which would enable individuals to challenge automated decisions made about them.[84] In particular, they argue that people should have a 'right to inspect, correct, and dispute inaccurate data and to know the sources (furnishers) of the data.'[85] Furthermore, they argue that an algorithm that generates a score from this data needs to be publicly accessible – rather than secret – so that each process can be inspected. Finally, they emphasise that policymakers need to ensure that a score is fair, accurate, and replicable.[86]

Where full transparency is not possible and is reasonably overtaken by other considerations, the accountability of the decision-making process as a whole still needs to be ensured. Qualified transparency can play a role – even a complex machine learning system can be evaluated and tested so that the impact of particular variables on outputs is measured. Differential impact on traditionally marginalised groups should be something that is tested before implementing an automated system, and sufficient access to the system should be faciliated to enable further testing.[87] However, where automated components of systems cannot be made transparent, accountability needs to be assured by humans. Ensuring a human is responsible for independently justifying the decision and that humans are involved in

[81] For example, the Artificial Intelligence, Ethics and Society (AIES) conference, the IEEE's (Institute of Electrical and Electronics Engineers) Global Initiative on Ethics of Autonomous and Intelligent Systems, the International Standards Organisation's JTC1/SC42 standardisation program, and the 'Artificial Intelligence Roadmap and Ethics Framework' project at Australia's Data61,

[82] A. M. Carlson, 'The Need for Transparency in the Age of Predictive Sentencing Algorithms' (2017) 103 *Iowa Law Review* 303; N. Diakopoulos, 'We Need to Know the Algorithms the Government Uses to Make Important Decisions About Us' *The Conversation* 24 May 2016 at http://theconversation.com/we-need-to-know-the-algorithms-the-government-uses-to-make-important-decisions-about-us-57869 (last accessed 16 August 2018).

[83] S. B. Starr, 'Evidence-Based Sentencing and the Scientific Rationalization of Discrimination' (2014) 66 *Stanford Law Review* 803, 806.

[84] D. K. Citron and F. Pasquale, 'The Scored Society: Due Process for Automated Predictions' (2014) 90 *Washington University Law Review* 1, 20.

[85] *ibid*.

[86] *ibid*, 22.

[87] Kroll et al, n 79 above. See also Citron and Pasquale, n 84 above, 25.

appeal processes, as is the case in Sweden, is one way in which accountability can be preserved. In these situations, it will be important to ensure that such humans feel able to act independently of the outputs of the automated system. Finally, it may be the case that, because of the inherent opacity, certain decision-making by the governments should not be delegated to software with particular characteristics. For example, to remain in line with transparency and accountability values that form part of the rule of law, criminal sentencing should not be fully or partially delegated to a system whose logic cannot be rendered transparent and comprehensible to defendants and their representatives. This ensures that factors that ought to be irrelevant in the sentencing process remain so.

Predictability and consistency

Automation can also improve the predictability and consistency of government decision-making. Unlike humans, computer systems cannot act with wanton disregard for the rules with which they are programmed. They can be programmed to act probabilistically, tossing a virtual coin to decide whether a decision is made in an applicant's favour, but such deliberate arbitrariness does not arise in any of our examples. Instead, it is generally reserved for situations where a social consensus supports randomisation as the only fair means of allocation (as with issuing limited tickets for an event or determining lottery winners).

As such, the systems in our examples generally enhance the predictability and consistency of decision-making, even where they are otherwise problematic. The social credit system in China works as a tool of social control *because* people can predict the consequences of engaging in particular activities that the government wishes to discourage. Australia's robo-debt program and Sweden's social welfare system perform the same calculation for everyone.

However, automation also poses many challenges for the rule of law principles of predictability and consistency. A first challenge arises when the rule that is applied in an automated decision-making process does not correspond with statutory or common law requirements. The inconsistency in such cases is not in the application of the rule in different cases, but between the rule as formulated and the rule as applied in every case. An example of such inconsistency is robo-debt. The formula failed to produce the legally correct result for many people.[88] This is not necessarily a problem where people are given the opportunity to correct matters, as is evident from the Department's defence of its position:

> Initial notices request information to explain differences in earned income between the Australian Taxation Office and Centrelink records. These result in a debt in 80 per cent of cases. The remaining 20 per cent are instances where people have explained the difference and don't owe any money following assessment of this updated information. This is how the system is designed to work, in line with the legal requirements of welfare recipients to report all changes in circumstances and the department's obligation to protect government outlay.[89]

The problem was not that there was an error rate, which also exists for decisions made by humans, but that the processes in place to manage the error were insufficient. There was no human checking of the decision to issue a debt notice. The notice itself was also presented to individuals as a fait accompli, with some individuals not receiving earlier communications due to address errors.[90] The online portal in place to deal with challenges to debt notices was

[88] There is some dispute about the rate of error and how these should be characterised. Approximately 20 per cent of people who received debt notices succeeded in providing additional information that demonstrated that no debt was owed: Senate Community Affairs References Committee, Parliament of Australia, *Design, Scope, Cost-Benefit Analysis, Contracts Awarded and Implementation Associated with the Better Management of the Social Welfare System Initiative* (2017) at [2.88].

[89] *ibid* at [2.89].

[90] *ibid* at [3.61].

also hard to use,[91] with human alternatives inadequate to meet the demand.[92] The rate of errors also potentially exceeded the capacity of institutions designed to deal with appeals. This compares unfavourably with the automated Swedish system, where humans edit and take responsibility for each decision, with usual processes in place for appeal.[93] The result in Australia is a far higher likelihood that the law is being misapplied in ways that are unpredictable and inconsistent.

When moving from pre-programmed rules to rules derived from data (for example, through supervised machine learning), the predictability and consistency of decision-making may be reduced. This is not because the computers are acting contrary to programming but because, like human children who 'learn', it is hard to predict the outcomes in advance and behaviour will change as 'learning' continues. Consider what is known about the COMPAS tool (which is limited due to the transparency issues discussed above). Those developing the tool did not necessarily know in advance what criteria would be found to correlate, alone or in combination, with particular behaviours (such as reoffending). The rules allocating scores to individuals were derived, likely through a supervised machine learning process, from a large set of data (namely data recording historic re-offending behaviour). The behaviour of the system is thus difficult, and sometimes impossible, for a human to predict in advance.

Machine learning raises another issue for predictability and consistency because it continues to 'learn' from new data fed into it over time. If it gives a low score to an individual, thereby contributing to a decision to grant parole, but the individual reoffends, that will be fed back into the algorithm in order to improve its predictive accuracy over time. In that way, a new individual who was relevantly 'like' the earlier false negative will have a different outcome, namely a higher risk score and lower chance of parole. This means that the system treats identitically situated individuals differently over time which, as discussed below, is a problem not only for consistency but also for equality before the law.

Further, there are differences in how judges and risk assessment tools assess the risk of re-offending. While the information that judges can consider for sentencing is not generally restricted by traditional evidentiary rules and can include factors about defendants' personal and criminal history,[94] the process of sentencing itself must satisfy the natural justice or due process requirements.[95] Accordingly, judges are unlikely to make sentencing decisions that hinge on inherent characteristics of defendants, such as whether their parents are divorced.[96] The fact that COMPAS relies on variables that would not have been considered relevant by a human judge creates an inconsistency between decisions made by judges under the law and decisions suggested by algorithmic inferencing. The lack of transparency about the data relied on in the machine learning process in a particular case, as

---

[91] *ibid* at [2.110].

[92] *ibid* at [3.98], [3.106], [3.107], [3.119].

[93] CSN decisions can be appealed to the National Board of Appeal for Student Aid (Överklagandenämnden för studiestöd, 'OKS'), see OKS website at https://oks.se/ (last accessed 6 November 2018).

[94] In the US, the Federal Rules of Evidence (FRE) generally do not apply at sentencing, see, eg, Federal Rules of Evidence r 1101(d)(3) (2015) at http://federalevidence.com/rules-of-evidence#Rule1101 (last accessed 10 September 2018). For a detailed discussion of US sentencing, see D. Young, 'Fact-Finding at Federal Sentencing: Why the Guidelines Should Meet the Rules' (1993) 79 *Cornell Law Review* 299.

[95] In the US, this has been recognised by the US Supreme Court in *Gardner* v *Florida* 430 U.S. 349, 359 (1977) (noting that '[t]he defendant has a legitimate interest in the character of the procedure which leads to the imposition of sentence even if he may have no right to object to a particular result of the sentencing process.').

[96] J. Angwin, 'Sample-COMPAS-Risk-Assessment-COMPAS-"CORE"' at https://www.documentcloud.org/documents/2702103-Sample-Risk-Assessment-COMPAS-CORE.html (last accessed 16 August 2018), showing the question 'If you lived with both parents and they later separated, how old were you at the time?'

well as opacity of the algorithm itself, makes it more difficult for judges to adjust their expectations of the tool to ensure appropriate use.

Automation can improve the predictability and consistency of decision-making by removing the arbitrariness for which humans are well known. However, the benefits can only be realised if the automation process is sufficiently transparent, if it is property evaluated (for accuracy and for consistency with legal requirements), and if appropriate measures are put in place to manage foreseeable errors. Such measures should include human checking of outputs, clear explanation as to the potential for error and the circumstances in which error can arise, and a transparent and sufficiently resourced process for appeals. These are all questions of design.

Automation according to human-crafted rules (derived from statute or judge-made law) can ensure that the correct decision is made every time and can overcome issues with human error and corruption. Rules derived from data raises more complex challenges, particularly in ensuring predictability and consistency with the 'law on the books.' Supervised machine learning and other iterative systems also struggle with consistency over time. However, these are matters that can be controlled from the perspective of predictability and consistency, in the first case through design of the system as well as independent testing and evaluation, and in the latter by moderating continual learning. Hence, a system that combines both types of automation by using explicit programming to automate the application of a fixed rule (originally derived from data, for example through machine learning) can ensure consistency over time. Automation *can* thus prove beneficial for predictability and consistency, although the evidence suggests that may not be achieved in practice.

Equality before the law

Automation can enhance the principle of equality before the law by reducing arbitrariness in the application of law, removing bias and eliminating corruption. For instance, automation in China's social credit system could, through the use of cameras and face recognition technology, be deployed to ensure consequences apply to everyone who breaches certain rules (such as jaywalking or parking illegally) without exception. By contrast, without such automation, systems in place for minor infringements of this kind require a person to be 'caught', with the severity of the penalty often depending on the discretion and 'generosity' of the officials in question. Moreover, the enhanced consistency discussed above, particularly of the expert systems, such as the Swedish welfare or robo-debt, that give the same answer when presented with the same inputs, helps to ensure that similarly situated individuals are treated equally. These examples demonstrate how certain kinds of automation can remove the capacity for biased humans to discriminate against unfavoured groups. A properly designed system could do so by eliminating both conscious and unconscious bias by only applying criteria that are truly relevant to making the decision.

The benefits that automation can provide to equality before the law are however qualified by two main interrelated challenges. First, automation in government decision-making might compromise due process rights and the extent to which the laws apply to all equally; and second, it might undermine the extent to which people, irrespective of their status, have equal access to rights in the law.[97]

[97] Interestingly, these concerns were also raised during the so-called selective incapacitation movement in the 1980s. Incapacitation theory sought to reduce crime rates by making offenders incapable of re-offending. As D. L. Kehl, P. Guo and S. A. Kessler explain, 'Selective incapacitation theory was based on the premise that the justice system should seek to identify, or "select," a sub-set of individuals who are particularly prone to violence

Firstly, automation can compromise individual due process rights because it may undermine the ability of that person to influence or challenge a decision affecting them. This may be, for example, because they are unable to access or determine the correctness of key information used to make that decision. For instance, in robo-debt, the right to review and rectify information was undermined because the letter sent to individuals by the government did not explain the importance of the income variation over the year for an accurate calculation of welfare entitlements.[98] Issuing debt notices for money not owed to a subgroup of welfare recipients without providing them with a genuine opportunity to correct the erroneous data held on them effectively denied them due process rights and hence equal treatment under the law.

By contrast, the involvement of a case officer in the Swedish student welfare example enables explanation of the process and provides an immediate opportunity for those affected to rectify information or exercise a right of review. Moreover, the process is strengthened by a relatively straightforward appeal procedure to challenge the CSN decisions.[99] For example, a student who had been prevented from joining the job market due to their disability had the initial CSN decision reversed after examination by the Swedish National Board of Appeal for Student Aid.[100] Decisions by the Board which are deemed to be of fundamental importance and in the public interest are available on its website.[101]

Similarly, under Shanghai Municipality SCS model, individuals have a right to know about the collection and use of their social credit information and can access and challenge the information contained in their credit reports.[102] The municipal Public Credit Information services centre will determine whether to rectify the information within five working days of receiving the objection materials. These rights were tested in practice by Chinese citizen Liu Hu, who was blacklisted on the SCS and unable to book a plane ticket after he accidentally transferred the payment for a fine to a wrong account.[103] After a court learned that Liu Hu had made an honest mistake, the information on his social credit report was rectified.

In the case of machine learning employed in government decision-making, lack of transparency, which is common for the reasons discussed above, is the primary reason why due process rights are compromised. In *Loomis*, the Supreme Court of Wisconsin held that

---

or recidivism—colloquially known as "career criminals"—and incapacitate them by keeping them in prison for longer periods of time': 'Algorithms in the Criminal Justice System: Assessing the Use of Risk Assessments in Sentencing' Harvard Law School Student Paper, Responsive Communities Initiative, Berkman Klein Center for Internet & Society, July 2017, 3 at http://nrs.harvard.edu/urn-3:HUL.InstRepos:33746041 (last accessed 16 August 2018). See also 'Selective Incapacitation: Reducing Crime Through Predictions of Recidivism' (1982) 96 *Harvard Law Review* 511; T. Mathiesen, 'Selective Incapacitation Revisited' (1998) 22 *Law and Human Behaviour* 455.

[98] Kehl, Guo and Kessler, *ibid*, 9.

[99] CSN decisions can be appealed to the National Board of Appeal for Student Aid (Överklagandenämnden för studiestöd, 'OKS'), see OKS website, n 93 above.

[100] The Swedish National Board of Appeal for Student Aid, Dnr: 2014-03172, available at https://oks.se/wp-content/uploads/2016/03/2014-03172.pdf (last accessed 6 November 2018).

[101] https://oks.se/avgoranden/ (last accessed 6 November 2018).

[102] 上海市社会信用条例 [Shanghai Social Credit Regulations] (Shanghai Development and Reform Commission, 29 June 2017) art 34 at https://www.chinalawtranslate.com/上海市社会信用条例/?lang=en (translated from http://www.shdrc.gov.cn/gk/xxgkml/zcwj/zgjjl/27789.htm) (last accessed 10 September 2018). Article 36 further states, 'Where information subjects feel that there was error, omissions, and other such circumstances … they may submit an objection to the municipal Public Credit Information service center, credit service establishments, and so forth.'

[103] S. Mistreanu, 'Life Inside China's Social Credit Laboratory' *Foreign Policy* (online), 3 April 2018 at https://foreignpolicy.com/2018/04/03/life-inside-chinas-social-credit-laboratory/ (last accessed 10 September 2018).

due process was preserved because a COMPAS score was only one among many other factors to be considered by the judge.[104] However, it is difficult to determine how far along the spectrum of automation a judge's use of this system will lie: in particular, some judges may treat it as a minor input into the decision, while others may be afraid that overriding the 'objective' evidence of dangerousness based on other considerations would be subject to public, political or appellate critique. Thus, the extent to which an individual decision is based on the outputs of COMPAS is difficult to assess.[105] Furthermore, there are reasons to believe that the score will have a greater influence than it deserves – the praise for such systems offered by institutions such as Conference of Chief Justices suggests the attraction of 'objectivity' has blinded many in the judiciary to the practical flaws of the software.

The Court in *Loomis* also added that the right to review and rectify was satisfied because the defendant had a degree of control over relevant input data: he could review the accuracy of public records and offer other data directly through completion of the COMPAS questionnaire.[106] However, there is a difference between the ability to review and rectify separate pieces of information which are fed into the software and the ability to review how the score is calculated. While the opportunity to input data may be an improvement on the Robo-debt system, this argument ignores the fact that the rules applied by the COMPAS system are derived from historic data and that none of the data, the machine learning technique, or the derived rules have been made public. The process through which a score is obtained is thus difficult to challenge. Further, a defendant lacks an effective opportunity to challenge the idea that factors outside of his control (for example, the fact that his parents divorced when he was three, asked in the COMPAS questionnaire[107]) influence the length of his sentence. Indeed, it would be impossible for a defendant to even know whether such a factor did influence his score, as the lack of transparency prevents a defendant from knowing the extent to which any given data (in public records or the questionnaire) has proved to be material. A defendant is therefore only given an opportunity to argue against a score in the absence of any real understanding of the basis for its calculation. Similar due process concerns because of lack of transparency also arise in parts of the SCS system.

Further challenges to equality before the law and due process safeguards can arise in some cases of automated decision-making due to what could be described as a 'reversal' of the burden of proof or lowering of the 'evidence threshold'.[108] For example, in the robo-debt

---

[104] It is likely significant that the judge told Loomis at the sentencing hearing that the COMPAS score was one of multiple factors that his Honour weighed when ruling out probation and assigning a six-year prison term: 'In terms of weighing the various factors, I'm ruling out probation because of the seriousness of the crime and because your history, your history on supervision, and the risk assessment tools that have been utilised, suggest that you're extremely high risk to re-offend', *Loomis* n 49 above, 755.

[105] The Court simply added that while COMPAS cannot be determinative in sentencing decisions, the risk scores can be considered a *relevant* factor in several circumstances, including: '(1) diverting low-risk prison-bound offenders to a non-prison alternative'; (2) assessing the public safety risk an offender poses and whether they can be safely and effectively supervised in the community rather than in prison; and (3) to inform decisions about the terms and conditions of probation and supervision, see *Loomis ibid*, 767–772 *per* Bradley, J, 772 *per Rogennsack*, CJ, concurring, 774 *per* Abrahamson, J, concurring.

[106] *ibid*, 765.

[107] See n 96 above.

[108] On the importance of burden of proof and 'evidence threshold' in the context of social welfare in the US, see L. Kaplow, 'Burden of Proof' (2012) 121 *Yale Law Journal* 738. For Australia, see, eg, A. Gray, 'Constitutionally Protecting the Presumption of Innocence' (2012) 31 *University of Tasmania Law Review* 13. In the context of European Court of Human Rights, see M. Ambrus, 'The European Court of Human Rights and Standards of Proof: An Evidential Approach Toward the Margin of Appreciation' in L. Gruszczynski and W. Werner (eds), *Deference in International Courts and Tribunals: Standard of Review and Margin of Appreciation* (Oxford: OUP, 2014). On due process implications of shifting the burden of proof in the US legal context, see C. M. A. McCauliff, 'Burdens of Proof: Degrees of Belief, Quanta of Evidence, or Constitutional

case, debt notices were issued for money that was not in fact owed by some welfare recipients, and the fact-finding burden for debt that previously rested on the Department was reversed, arguably contrary to the enabling legislation.[109] While debts issued under this automated decision-making process can be challenged, it has been argued that the government failed in its responsibility to ensure that it has established the existence of the debt before initiating the claim.[110]

Finally, the use of automated decision-making by the governments poses a further challenge to the idea that all individuals irrespective of their status must have equal access to rights in the law, and that in accessing these rights 'like cases be treated alike'. This includes the notion that government should not treat individuals differently due to their demographic group or an immutable trait.[111] Automated decision-making systems, such as COMPAS and Sesame Credit can undermine this principle because they may: a) explicitly incorporate and rely on various static factors and/or immutable characteristics, such as socio-economic status, employment and education, postal codes, age or gender; or b) take such matters into account indirectly, for example by 'learning' the relevance of variables that correlate with these. For example, in *Loomis*, the defendant has argued that the judge's consideration of the COMPAS score also violated his constitutional rights because COMPAS software used 'gendered assessments',[112] and in in turn undermined his right to an individual sentence. As was mentioned in the previous section, the use of COMPAS and similar sentencing software, might permit judges to apply factors and characteristics that have long been considered inappropriate in the context of criminal sentencing.

The greatest challenge to equality before the law comes not from an explicit incorporation of inappropriate variables in the automated system, but from the fact that automation can infer rules from historical patterns and correlations. Even when variables, such as race, are not used in the learning process, a machine can still produce racially or otherwise biased assessments. As was mentioned earlier, a 2016 ProPublica investigation found that African Americans are more likely than whites to be given a false positive score by COMPAS risk assessment software, despite the (claimed) fact that race is not used as a variable.[113] This unequal treatment before the law results because many other factors can correlate with race, including publicly available information, such as, eg, Facebook 'likes' which are not excluded from the machine learning process.[114] Further, the data from a pre-sentencing questionnaire (from which the COMPAS tool draws inferences) records the

---

Guarantees' (1982) 35 *Vanderbilt Law Review* 1293; P. Petrou, 'Due Process Implications of Shifting the Burden of Proof in Forfeiture Proceedings Arising out of Illegal Drug Transactions' [1984] *Duke Law Journal* 822.

[109] Hanks, n 46 above.

[110] Carney, n 45 above.

[111] People have particularly strongly objected to courts systematically imposing more severe sentences on defendants who are poor or uneducated or from a certain demographic group: see G. Kleck, 'Racial Discrimination in Criminal Sentencing: A Critical Evaluation of the Evidence with Additional Evidence on the Death Penalty' (1981) 46 *American Sociological Review* 783; L. Wacquant, 'The Penalisation of Poverty and the Rise of Neo-Liberalism' (2001) 9 *European Journal on Criminal Policy and Research* 401; C. Hsieh and M. D. Pugh, 'Poverty, Income Inequality, and Violent Crime: A Meta-Analysis of Recent Aggregate Data Studies' (1993) 18 *Criminal Justice Review* 182.

[112] *Loomis*, n 49 above, 757.

[113] Angwin et al, n 52 above.

[114] See especially, M. Kosinski, D. Stillwell and T. Graepel, 'Private Traits and Attributes are Predictable from Digital Records of Human Behavior' (2013) 110 *Proceedings of the National Academy of Sciences of the United States of America* 5802 (finding that easily accessible digital records such as Facebook 'likes' can be used to automatically and accurately predict highly sensitive personal information, including sexuality and ethnicity).

number of times and the first time a defendant has been 'stopped' by police. Given historical profiling practices of law enforcement in the United States, status as an African-American is likely to correlate with higher numbers and earlier ages in response to this question.[115] Racial differentiation is thus built into the data from which correlations are deduced and inferences drawn.

Unlike the risk assessment tool COMPAS, decisions in Swedish student welfare management system are made solely on factors that are legally relevant. The pre-programmed nature of the system ensures that those factors play a role in the decision *precisely* in the circumstances in which they are relevant. Decisions are made consistently with the law, with students treated equally under that law. In Chinese SCS, diversity of implementation means that equality before the law is affected differently. For example, decisions in the Rongcheng City model of the SCS system are made solely with reference to clearly defined categories of behaviour which leads to either a point deduction or addition – there is no room to consider any other factors in the pre-programmed system. In contrast, however, the Sesame Credit system in the SCS relies on variables that are irrelevant from a rule of law perspective, such as the rankings of an individual's social network contacts, which could lead to differential treatment in effect based on social status, sex or ethnic origin.[116]

As our examples demonstrate, in understanding the benefits and challenges of automating government decisions, it is crucial to consider both the context of the decision and the type of system deployed. A system with pre-programmed rules can ensure that decisions are made based on factors recognised as legally relevant and hence avoid or minimise the risk of corruption or favouritism by officials. However, procedural rights and opportunities to check and rectify data on which the decision relies are crucial, as is ensuring that the logic of the system accurately reflects the law. As our case studies demonstrate, the challenges posed by systems based on rules inferred from data are different. Here, the role of humans is limited to setting parameters, selecting data (possibly biased due to flawed human collection practices), and deciding which variables to use as a basis for analysis. Unless the humans involved in these processes have a deep understanding of the legal context in which a decision is made, systems may fail in practice to meet the standard of equality before the law. The COMPAS system is an example of software that does not meet the needs of a fair criminal justice process – lack of transparency in a tool that relies on a large set of often legally irrelevant inputs prevents a defendant from having sufficient opportunity to participate in the court's findings on dangerousness, which is crucial component of the ultimate decision. The fact that the tool has more 'false positives' in the African-American community than among white people is further evidence that humans are exercising insufficient control over the machine learning process to ensure that it operates appropriately.

This does not imply that systems that rely on rules derived from data, including those deploying machine learning techniques, can never be used in government decision-making in ways that do comply with equality before the law. Machine learning can be used in the development of high-level policies, from traffic flow management to modelling interventions in the economy. Even at the level of decisions affecting individuals, machine learning is sometimes consistent with or even of benefit to equality before the law. Facial recognition, if designed to recognise the faces of diverse individuals accurately, could be used to identify individuals where that is an aspect of the system, and if programmed correctly may even overcome conscious and unconscious bias on the part of humans. While concerns about

[115] C. O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (New York, NY: Broadway Books, 2016) 25–26 ('So if early "involvement" with the police signals recidivism, poor people and racial minorities look far riskier.')
[116] See Kosinski et al, n 114 above.

privacy and surveillance may counter its benefits, the use of machine learning in such a system can improve equality before the law by reducing arbitrariness.

CONCLUSION

Automation can improve government decision-making. The benefits include cost savings and greater speed, as well as a capacity to enhance the rule of law. Properly designed, implemented and supervised automation, whether in the form of systems applying pre-programmed rules, systems that learn rules from historic data, or combinations of these, can help government decision-making better reflect the values of transparency and accountability, predictability and consistency and equality before the law.

What is apparent, though, is that three of the four studies of automation considered in this article fail to live up to this ideal. In some cases, such as robo-debt, this failure results from poor design and implementation of the automated system. Indeed, one consistent theme is that human choices, and often error, at the design and implementation stage of automation can cause a system to fail to meet rule of law standards. A contrast is the Swedish student welfare system, which involves high levels of automation, but does not raise the same concerns. The Swedish model, which puts a strong emphasis on compliance with national legislation, officers' ethical codes, and publishing of the rules, demonstrates how a carefully designed system integrating automation with human responsibility can realise many benefits, while remaining sensitive to the values expressed in the rule of law.

It would nonetheless be a mistake to suggest that effective human design and implementation can ensure a particular automation technique will enhance or at least meet the minimum standards of the rule of law. It is clear from our study that even with active human engagement some forms of technology raise intractable problems. This may be because the form of automation is inappropriate for its context. For example, machine learning offers many benefits, but some techniques or software products come at the price of transparency, and so accountability. This may be tolerable in particular circumstances, such as in the distribution of low-level welfare benefits (with appeal mechanisms), assisting with tasks such as optimising the traffic flow in a city, or conducting facial recognition for identification purposes. In such cases, testing and evaluation of accuracy and disparate impact may be sufficient from the rule of law perspective.

On the other hand, machine learning that cannot be rendered transparent and comprehensible may not be appropriate where it is used to make decisions that have greater effects upon the lives and liberty of individuals. It can also be inappropriate where a machine learning system may be influenced by criteria that ought not to be relevant, such as a person's race or even variables that have not traditionally been used to discriminate, such as the credit rating of one's friends. Such problems are exacerbated, as in the case of COMPAS, when the system operates according to undisclosed, proprietary algorithms. These problems would be compounded if COMPAS were used not only to assist judges, but to replace them.

From the perspective of the rule of law, these problems may become more acute over time. As technology develops, and machine learning becomes more sophisticated, forms of automation used by government may increasingly become intelligible only to those with the highest level of technical expertise. The result may be government decision-making operating according to systems that are so complex that they are beyond the understanding of those affected by the decisions. This raises further questions about the capacity of voters in democratic systems to evaluate and so hold to account their governments, including in respect of compliance with rule of law values. Ignorance in the face of extreme complexity may enable officials to transfer blame to automated systems, whether or not this is deserved.

The result may be an increasing tension between automation and the rule of law, even where humans design systems in ways that seek to respect such values.

Ultimately, humans must evaluate each decision-making process and consider what forms of automation are useful, appropriate and consistent with the rule of law. The design, implementation and evaluation of any automated components, as well as the entire decision-making process including human elements, should be consistent with such values. It remains to be seen whether these values can be fully integrated into automated decision-making and decision-support systems used by government. Converting rule of law values into design specifications that can be understood by system designers, and enforced through regulation, professional standards, contracts, courts or other mechanisms represents a formidable technical and legal challenge. This article highlights a number of common themes in this respect, including the need for an awareness of the link between tools/design and transparency/accountability, the need to consider consistency and predictability not only over time but also as between automated and human systems, the importance of embedding procedural due process rights, and the tension between deriving rules from historic data and equality before the law. Resolving these issues in the automation of government decisions will be critical for any nation that claims to uphold the basic ideals of the rule of law.

A deeper question beyond the scope of this article is the extent to which automation of government decision-making will itself shape the rule of law. The rule of law is not a static concept. It evolves in response to changing societal values and the operation of government. As technology reshapes society, and government interacts with the community, it can be expected in turn that our understanding of the rule of law will shift. Values such as transparency and accountability, predictability and consistency and equality before the law may remain central to conceptions of the rule of law, but their interpretation and application may change. The benefits offered by such technologies, such as their capacity to reduce government spending, may be so significant as to demand greater accommodation within the rule of law framework.