Legal Audit of AI in the Public Sector

HMI Technology Policy Paper: 1/2021

Will Bateman (Australian National University, Law School)

Julia Powles (University of Western Australia, Law School)

7 May 2021

ANU Humanising Machine Intelligence Grand Challenge

The authors acknowledge the generous support of the Minderoo Foundation.

Table of Contents

Executive Summary

1. The collection of technologies which are captured under the umbrella phrase 'artificial intelligence' (**AI**) trigger a re-think of the law applying to governments because they fundamentally change the power balance between public officials and citizens.

2. Automation, machine learning, data archiving/networking and mass surveillance technologies give the entities which control their use enormous advantages over the people who are subject to them.

3. Existing legal (and constitutional) frameworks applying to government are built on a 'human-centric assumption': that the people who exercise public power have the same cognitive, physical and social capacities as the citizens they govern. That assumption no longer holds when governments apply AI technologies that are more powerful, yet potentially more opaque and narrow-minded than human decision-makers.

4. Generally-speaking, legal rules that currently apply to government use of AI:

    a. lag behind technical advancements in AI;

    b. fail to explicitly regulate the potential harms of AI; and

    c. use 'soft' rather than 'hard' law.

5. More detailed conclusions can be reached about the law governing public sector AI by analysing case studies that show the way existing legal rules operate in concrete contexts.

6. To undertake that task, we use basic requirements of liberal democratic government as criteria to measure the appropriateness of existing legal frameworks applying to government use of AI (**Audit Criteria**):

    a. **Knowledge** of the essential features of how AI technologies use information and reach outcomes in a particular context;

    b. **Assent** to the use of AI through specific authorising legislation;

    c. **Personhood**, or treating people as autonomous individuals, as the basic standard for legitimate government behaviour;

    d. **Protection** of basic civil rights;

    e. **Contestability** before an independent judicial body; and

    f. **Remedial action** for wrongful use of AI.

7. We use those criteria to audit case studies which show how the law works ‘in application’ rather than ‘in theory’:

   a. Automation of welfare state functions in Australia via the Online Compliance Intervention (**OCI** or **robodebt**) system;

   b. Data-driven machine learning technology as part of the criminal justice system in the US via the Correctional Offender Management Profiling for Alternative Sanctions (**COMPAS**) system;

   c. Data archiving/networking in the UK National Health Service, which led to the sharing of personal health information by public health authorities with Google (**NHS/Deep Mind**); and

   d. Mass surveillance in UK policing via the use of live facial recognition technology (**NeoFace Watch**).

8. In each case study we assign a score to the legal frameworks governing AI by reference to the Audit Criteria, and reflect on how that score could be impacted under different legal regimes applying to AI.

9. We conclude the Legal Audit by presenting an assessment of the success of existing law in governing the use of AI by government.

Part I: Technological Realities

All reform-oriented analyses of the law and artificial intelligence must grapple with two critical questions:

1. What is artificial intelligence?
2. Informed by what this technology can/cannot do, should artificial intelligence trigger a re-think of existing legal frameworks; in this case, legal frameworks applying to governments?

This Part of the Legal Audit addresses those questions, before moving to introduce the legal landscape governing AI in the public sector (Part II) and the more detailed audit of the way law governs different types of artificial intelligence (Part III).

What is artificial intelligence?

The first task of this Audit is to provide some clarity on the technologies which answer the description **artificial intelligence (AI)**.

Although the expression 'AI' is ubiquitous, it is a chameleon. For some, AI only means 'machine learning'; for others, it appears to mean *any* sophisticated use of computerised processes. Each of these contemporary uses fall short of the classic definition of AI, being the creation of 'machine intelligence'. Given how quickly technology (and its applications) can change, a degree of vagueness in the meaning of AI is not surprising.

For the purposes of this Audit, we confine our definition of AI to the following four types of technologies:

- Automation;
- Machine learning;
- Data archiving/networking; and
- Mass surveillance.

As we explain in this Part, there is overlap between each of those technologies, but dividing them permits a clearer analysis of the distinct/similar issues which arise in relation to different technologies. Ultimately, each technology attempts to simulate aspects of basic human cognitive and physical processes, but in each case they produce effects, both positive and negative, on people and societies which require a fundamental re-think of the legal frameworks which apply to government.

Automation

The most basic AI technology is software that automates processes via deterministic code. Deterministic code involves conditional rules, driven by yes/no responses. This is often called 'rule-based automation' in the computer science community. It is referred to simply as **automation** in this Audit.

Automation is best understood as a type of advanced calculator which works by applying logical rules to inputs supplied by a user in order to produce outputs. At its simplest:

if input = X → produce output Y, otherwise → produce output Z.

Automation technology augments the capacity of individual humans to reason logically and deductively. To take an example:

- ➢ I must give a social security benefit to a person who:
  - Is under 45 years of age;
  - Earned less than $10,000 in the last tax year; and
  - Has no criminal convictions.
- ➢ Person A is 25 years old, earned $5,000 in the last tax year and has no criminal convictions.
  - I must give Person A a social security benefit.
- ➢ Person B is 25 years old, earned $10,500 in the last tax year and has a pending court date for a driving offence.
  - I must not give Person B a social security benefit.

The advantage of automation is that it permits clear rules to be applied in pre-determined ways without the intervention of cognitive and physical limitations. Compared to an individual human thinker, computers running rule-based automation never get tired, emotional, hungry or bored. Because automated processes lack those human limitations, they will never misapply a simple rule or miscalculate a figure. However, automation technologies are also significantly-limited in comparison to an individual human thinker, in that they are *only* capable of applying sets of nested simple rules and making calculations.

Automation technology already has a wide application in the public sector and can be employed in virtually any citizen-facing or departmental domain. Some high-profile applications include:

- ➢ Taxation (performing taxation assessments and responding to simple taxpayer appeals);
- ➢ Social welfare (determining whether social insurance benefits should be paid);
- ➢ Public health (determining whether health care subsidies should be paid);

- Civil enforcement/policing (responding to simple internal appeals against fines); and
- Immigration (granting visas/entry permits).

Automation technologies produce obvious economic, administrative and social benefits, including:

- Decreasing the cost/unit of goods and services (via overall reduced labour costs; though automation requires often less visible, distributed, low-wage labour to produce and maintain the technology);
- Increasing the output of goods and services/reducing economic slack (via reduced production/processing time);
- Increasing the accuracy of simple, repetitive tasks (via predictability and scalability); and
- Increasing social utility (through a combination of the above).

Automation technologies do, however, have the capacity to produce significant harms to individual and social groups, including:

- Oversights and errors resulting from the impossibility of reducing complex social practices to deterministic code and the patchiness of the information and conditional statements that inform automation systems;
- Oversights and errors resulting from the time-lag between social and environmental conditions at the time of coding and the time of application of an automated system; and
- Amplification of those errors as a result of the scale and speed of computing power.

In essence, real social harms can result from automated systems which are overly reductive and crude, incorrectly coded, or cannot be coded to respond to changing social and environmental realities.

Machine learning

The highest-profile AI technology is the use of machine learning models to analyse large quantities of information (expressed in data sets) in order to classify and recognise patterns in historical data, and then to use those patterns to make probabilistic predictions about future actions. We call that AI technology **machine learning (ML)** in this Audit.

Machine learning technologies are regarded as ‘data-driven’ because they require very large historical data sets. Those technologies allow ‘data-informed’ predictions, because they augment the capacity of humans to make predictions about the future based on historical information.

Consider the following familiar process of human reasoning:

- ➢ I need to hire a reliable employee.
  - I have hired employees in the past who do/do not have more than 2 years in continuous employment.
  - In general, those previous employees who held a job longer than 2 years were more reliable.
- ➢ Potential employee A has worked for 3 years in her past job, and so I predict she is reliable and will hire her.
- ➢ Potential employee B has never held a job for more than 8 months, so I predict that she is not going to be reliable and will not hire her.

Machine learning technologies do something akin to these basic reasoning processes to make probabilistic decisions from historical data sets. As with automation, machine learning systems have several advantages over human thinkers, but they also have significant shortcomings.

The first advantage of a machine learning system is granularity. Using the example of hiring a reliable employee, the use of a machine learning system may result in the following deliberative steps:

- ➢ The target of optimisation, i.e. what is a 'reliable employee', will have to be defined more precisely by the user of the system in order to train a machine learning model. For example, the user may define a reliable employee as an individual who:
  - (i) will not infringe a workplace policy; and
  - (ii) will remain in the job for more than 4 years.
- ➢ Based on a dataset of previous employees where (i) and (ii) are known, a machine learning model can make a prediction. For example, fed by historical data such a system may predict that:
  - Potential employee A is 20% more likely to be reliable than potential employee B.
- ➢ Additionally, if the data set is tagged to include demographic data on employees, this can be used to further refine and quantify the prediction of 'reliability'. Take as an example:
  - *Age*: Employee A (21); Employee B (40);
  - *Qualification*: Employee A (high school diploma); Employee B (doctorate); and
  - *Relationship status*: Employee A (single); Employee B (married with children).

    - A machine learning model might predict in these circumstances that employee A is only 3.876% more reliable than employee B.

- A machine learning model is also capable of indicating how initial input factors (e.g., age, qualification) contribute to the final classification result.

Like automation, machine learning systems have the advantage of being able to perform tasks at immense speed and accuracy compared to a human thinker (or group of thinkers). In that way, the principal social benefits of machine learning systems mirror those of automation:

- Increased economic value to users and increased economic output; and

- In theory, social gains based on accuracy and transparency of decision-making, if the information that drives machine learning outputs is accessible and understandable.

Also like automation, machine learning systems can produce very significant social harms.

Machine learning algorithms often make the wrong predictions based on a number of factors, including:

- *Imperfect historical data sets:* for example, where turns out that the system above was trained on a data set comprised of 1,000 people working in a car factory in Michigan in 1975, and did not represent the same 'reliability' attributes as people applying for jobs in a digital-marketing firm in 2020.

- *Insufficient information about prediction*: for example, all the jobs which employee A has worked may be positions with family members, while employee B may have worked a number of casual contracts for the same group of companies, removing any meaningful basis for differentiating them. Additionally, a recent graduate may have no meaningful employment history, incorrectly indicating the absence of reliability.

- *Inability to see into the future*: for example, the 'reliability' of both employees A and B might be materially impacted by changing factors which are external to the data set and the data inputs. Such factors could affect labour market competition (e.g., the availability and attractiveness of other jobs) and the economy more broadly (e.g., climate change motivating both employees to leave a carbon-intensive industry on moral grounds).

- *Mistakes about the relevance of a given prediction*: for example, focusing on the 'reliability' of an employee may be less relevant than the potential economic benefits of hiring that employee, even if they later prove to be unreliable.

The social harms caused by these deficiencies in machine learning have been the subject of extensive academic literature (for a recent survey of the (fast-developing) field, see: "A Survey on Bias and Fairness in Machine Learning"). In contrast, the social benefits of these technologies are frequently vaunted, despite the absence of any strong evidentiary basis about those benefits (see: PwC 2017; Eggers et al. 2017; EY Global 2018).

Data archiving/networking

Another potent AI technology at the foundation of both automation and machine learning are digital systems capable of archiving very large amounts of information and transmitting that information through digital networks. We label those technologies **data archiving/networking**.

Dramatic increases in computational capacity have enabled an explosion in the gathering, sharing and analysis of information about individual people's lives via machine-readable databases and networking infrastructure. Critically, much of that information is provided by citizens in the course of engaging with government service providers, rather than being collected surreptitiously (contrasting with 'mass surveillance technologies' explained below).

The essential features of data archiving/networking rely on:

- *Digitisation*: coding of information in machine-readable format;
- *Data archiving*: storage and ordering of information in large data sets; and
- *Digital networking*: connecting many digital computers to those data sets.

Like automation and machine learning, data archiving/networking has many applications in the public sector, including:

- Law enforcement (via databases of criminal conduct);
- Social welfare (via databases of welfare recipient behaviour);
- Health (via databases of hospital and public health agency records); and
- Immigration (via databases of entry/exit to/from a country).

Data archiving/networking has parallels and differences with the way humans gather and share information:

- While human cognitive and physical capacities limit the accuracy and scale of archived information through limits on human memory and archival space, data archiving/networking technologies never forget, get tired, or run out of space.
- Human networks have physical and social limits on their capacity to share information, including limits arising from strategic behaviour where people use information as an economic or social resource, and constrict its supply in order to boost the level of demand in people who desire the information. Data archiving/networking technologies are spared these physical limits, though

often replicate the social limits through the same forces of competition and protection.

- Human archives and networks are capable of preserving context and layering tacit and other forms of knowledge, while digital archiving/networking has poor retention of context and little capacity to support diverse forms of knowledge.

In common with automation and machine learning, there are efficiency arguments to be made in favour of data archiving/networking technologies. However, these same technologies have the potential to cause harm, including:

- Digital records are imperfect and whether or not data is created and archived can have multiple causes, but nevertheless always has political consequences (e.g., in Australia, there are more data archives on Indigenous Australians that reflect a deficit model than an empowerment model);
- Data archives/networks lack context, creating both individual and collective harms;
- The widespread collection and monetisation of people's private information and social behaviour without recognition for fundamental rights to privacy, freedom of association, and without distribution of economic benefits to the human originators of the data.

Mass surveillance

The final type of AI technology considered in this Audit appears in systems which permit the observation and recording of human activities on a massive scale. We label those technologies **mass surveillance**.

Mass surveillance technologies permit the instantaneous observation, recording and storage of information concerning individual human behaviour and human social interactions, including:

- Voice;
- Text (whether hardcopy or softcopy);
- Images (including facial recognition);
- Biometrics (biological information unique to a single human being, as well as inferred from populations); and
- Geolocative data.

While mass surveillance can integrate automation, machine learning and data archiving/networking, it essentially relies on hardware technology of the following kinds:

- Cameras;
- Biometric scanners: fingerprint, voice, retinal, facial, gait, body;

- Mobile computing devices and applications: laptop/desktop computers, tablets, smartphones;
- Physical access points for those cameras, scanners and devices; and
- Networks which permit transmission of information collected from those cameras, scanners and computing devices to storage facilities.

Like data archiving/networking, mass surveillance augments human information gathering processes by removing various biological and physical limitations, whether intellectual (forgetting faces/dates/times), biological (the physical or mental alertness of a biological observer); physical (physical sensory organs, including eyes, skin, ears, vestibular system, nose); or social (mis-recording or withholding information for strategic purposes, at least not in a manner which is non-transparent in the code).

Mass surveillance has many applications in government, including:

- National security (anti-terrorism and military uses);
- Policing and justice (preventative and investigatory policing and civil enforcement);
- Immigration (monitoring non-citizen groups);
- Public health (monitoring disease transmission); and
- Whole of government (formulation of social policy based on mass surveillance).

Mass surveillance is attractive to governments because of its capacity to boost the capacity of law enforcement and other agencies to enforce legal rules.

Despite that attractiveness, mass surveillance can produce very harmful effects on individual humans and social groups:

- A chilling in social and political expression, including artistic and intellectual activities;
- A loss of fundamental rights;
- An underinvestment in other, proven techniques for conducting investigations and ensuring security and safety;
- A transferral of economically valuable resources from surveilled citizens to governments or commercial third-party technology providers.

What does AI change about government?

A set of three factors combine to trigger a re-think of the legal frameworks applying to governments in light of AI.

1. Governments are effective (and desirable) because of their unique power over individual citizens, as a collective, and representative, of citizens: ie, power of the people;
2. Government power is expected to be informed, and limited, by human cognitive, physical and social features, and legal principles applying to government reflects those 'human-centric' expectations.
3. Advances in AI challenge human-centric assumptions of government and the legal frameworks applying to government.

Together, those factors warrant a re-examination of the basic legal relationships (or 'balance of power') between individual and state expressed through law.

What is 'government'?

A government is a concentration of social power: the ability of a large group of people (a majority of the electorate) to impose their decisions on other people (the electoral minority, members of the electoral majority who disagree with government decisions and non-citizen residents).

That concentration of power is often desirable.

The unique powers of governments can lift people out of poverty by ensuring a social minimum standard which provides goods and services which the market cannot deliver (welfare state policies). Governments can also permit widespread cooperation between people to achieve complex projects (such as transport, residential and recreational infrastructure) and protect the personal safety and life ambitions of individuals (through police, judges and armed forces).

Those positive outcomes of government only exist because the state has a monopoly on the legitimate use of force. Government can decide how people in a society should behave, and enforce those decisions through sanctions which reduce personal liberty (ultimately through police) or imposing economic constraints on them (through taxation and regulation of commercial activity). In that way, a fundamental power imbalance between individual citizens and public officials permits governments to achieve positive outcomes.

That imbalance of power can also produce negative outcomes.

Some are obvious, such as aggressive assertions of official power which cause unnecessary or disproportionate harm of individuals: for example, when police enforce the law using unnecessary violence or invasions of people's private lives which are unnecessary to reduce crime.

Some negative results of government power are less obvious. For example, governments can formulate or enforce policy based on incorrect factual information about a social cohort, or the impact that their interventions will have on that cohort. An example of that type of harm is the imposition of criminal liability on victims of domestic violence for continuing to have contact with abusive partners in violation of restraining orders, which follows from misunderstanding the psychological and physiological causes of family violence (For some of the complex issues that arise in that context, see: "Apprehended Violence Orders").

Controlling government through law

The various legal regimes which apply to government can be understood as a compromise between the positive and negative potential of the imbalance of power between citizens and public officials.

Constitutional law provides foundational authority for governmental institutions to exercise power in ways which are connected to democratic elections, supervised by judicial bodies which are independent of government and constrained by basic civil and political rights. Depending on the jurisdiction, constitutions may explicitly protect basic liberal rights, or those rights may be protected by judicial interpretation of more 'rights-neutral' constitutions.

Administrative law (the general law applying to government action) is designed to balance the need for efficient government with decision-making procedures which are reasonable, fair and transparent. The dignity of citizens and the efficiency of government decision-making are both protected by administrative law frameworks. These frameworks ensure procedural fairness, prohibit biased decision-making, staple bureaucratic action to parliamentary legislation and require public officials to provide transparent and rational reasons for their decisions.

Human rights law protects an irreducible core of civil and political rights from infringement by governments: freedom of association, conscience, speech, freedom from arbitrary detention, electoral rights and due process rights. Human rights law does not, however, impose unyielding limits on government power, but permits rights infringements which are 'proportionate' to humane policy objectives. In that way, human rights law acknowledges the positive and negative results of government power.

Anti-discrimination law is a type of applied human right, providing special protection of certain human characteristics from discrimination: sex, gender, race, religion, age, sexual orientation, physical ability. Privacy and data protection laws are also types of applied human rights protection, imposing obligations on governments (and private sector entities) to obtain consent before obtaining, storing and transferring people's personal information. Both anti-discrimination and privacy law do, however, permit governments limited rights to discriminate (through affirmative action regimes, and exemptions for certain public policies) and to gather/use personal information obtained without consent (for law enforcement, public health, social welfare and many other activities of government).

Human-centric expectations of government power and public law

Each of those legal frameworks is premised on an assumption that public officials and citizens have the same basic cognitive, physical and social capacities.

Constitutional and administrative law assume that the humans who will be elected to parliaments and exercise the powers of the executive government mirror, in all relevant respects, the intellectual and physical capacities of their fellow citizens.

Human rights law assumes that proportional trade-offs between human rights and other desirable social objectives will be achieved through the use of ordinary human cognitive processes. Anti-discrimination law assumes that governments' decision-making capacities will mirror those of citizens, and privacy law assumes that ordinary citizens can interact with one another without exposing their interactions and behaviour to automatic collection, transmission and analysis by AI technologies.

The augmentation and distortion of human cognitive, physical and social capacities which flows from automation, machine learning, data archiving/networking and mass surveillance technologies fatally undermines those assumptions.

Part II: Legal Landscape

The current legal landscape applying to the use of AI in the public sector is defined by three features:

1. Regulatory time-lag;
2. Piecemeal approach; and
3. A preference for 'soft' rather than 'hard' law.

Regulatory time-lag

Legal scholars and historians of technology have long noted the delayed reaction of legal regimes to technological advances (See: Moses 2013; Pasquale 2018). Consistent with this trend, and despite the widespread adoption of AI technologies in the public sector, no advanced economy has developed a comprehensive legal framework to govern the use of AI by government.

Adoption of automation and data archiving technologies began in the 1970s and 1980s within the OECD. Prominent examples included the deployment of automated tax administration software in the US from the 1970s, and welfare agencies in Northern Europe.[1] Throughout the 1990s, the deployment of automation throughout government was normalised, particularly in law enforcement and welfare agencies (See: "Technological Due Process"), without any root and branch re-thinking of the styles of legal regulation which applied to government action.

From the 2000s, the trend of replacing human decision-makers with AI technologies in government accelerated enormously. Automated systems were rolled out throughout government, and were matched with the use of machine learning, data archiving/ networking and mass surveillance technologies. From the 2010s, a series of high-profile revelations showed the world that the deployment of AI by governments was not going to be business-as-usual:

- The disclosures of mass surveillance by US and Five Eyes national security agencies in 2013 by Edward Snowden (See: "Surveillance, Snowden, and Big Data: Capacities, consequences, critique");
- Widespread racial discrimination in use of automated risk assessment tools used by US criminal justice authorities and courts revealed in 2016 (See: "Machine Bias";

[1] Westin, *Privacy and Freedom* (1970); Fuster, *The Emergence of Personal Data Protection as a Fundamental Right of the EU* (2014).

- The unlawful transfer of millions of confidential patient records by the UK National Health Service to Google in 2015 (See: "Google DeepMind and healthcare in an age of algorithms"); and

- The unlawful use of automated debt-recovery algorithms by the Australian government in 2016 (See: "The New Digital Future for Welfare: Debts Without Legal Proofs or Moral Authority?").

Despite the enormous controversy which attended those scandals, the basic assumption of parliaments and legislators was that existing legal frameworks were not in need of fundamental analysis or critique. Administrative law has not changed in response to the normalisation of automated decision-making by government agencies. Anti-discrimination law has not been amended or updated in order to respond to the rapid and diverse adoption of machine learning, data archiving/networking and mass surveillance technologies.

In that sense, the legal rules regarding AI have not kept up with advances in use of the technology. The one arguable exception are data protection laws that attempt to strictly regulate data flows in approximately 100 countries. The most updated of these regimes in the 2016 *General Data Protection Regulation* adopted by the European Union. As we explain below, that legal framework has a dubious claim to creating a comprehensive regulatory regime of the use of AI by governments. Data retention and investigatory powers are another area that has seen significant legislative activity, but without creating any comprehensive regime.

In one sense, the regulatory time lag of public law should come as no surprise, as many areas of law lag behind technical and social developments. The force multiplier which AI attaches to government powers does suggest, however, that significant concern should attend the legal time-lag over AI.

Piecemeal approach

The piecemeal quality of legal attempts to grapple with AI use by government is the second defining feature of the current legal landscape applying to the use of AI in the public sector.

We use the term 'piecemeal' to reflect two features of the legal frameworks:

1. *No comprehensive regulation*: where law applies specifically to AI (often captured imperfectly through legislative references to 'a computer', or 'software'), it is simply appended to existing legal frameworks which assume non-enhanced and non-degraded human cognition as the default.

2. *No meaningful regulation*: AI-specific laws tend to avoid meaningful regulation of technological subject-matter.

Despite AI being used in progressively expanded ways for over 40 years by governments, no dedicated legal instrument exists to regulate that use.

The most prominent examples of the existing piecemeal approach are the bare 'computerised decision-making' authorising provisions of legislation which exist in the UK and Australia.

In 1998, the UK parliament conferred decision-making authority on 'computers' under its major social welfare legislation:[2]

> *2.—(1) Any decision, determination or assessment falling to be made or certificate falling to be issued by the Secretary of State under or by virtue of a relevant enactment, or in relation to a war pension, may be made or issued not only by an officer of his acting under his authority but also–*
>
> *(a) by a computer for whose operation such an officer is responsible…*

In 2001, the Australian Parliament enacted a very similar law in its national social security legislation:[3]

> *(1) The Secretary may arrange for the use, under the Secretary's control, of computer programs for any purposes for which the Secretary may make decisions under the social security law.*
>
> *(2) A decision made by the operation of a computer program under an arrangement made under subsection (1) is taken to be a decision made by the Secretary.*

Those laws authorising the use of computers in administering social welfare department functions illustrates the piecemeal nature of legal responses to the use of AI by governments. They are replicated in a broad swathe of legislation in each of the UK and Australia governing, among other important topics, public education, citizenship, sovereign finance, biosecurity and immigration.

Such authorising legislation provides no comprehensive regulation of the subject-matter, but simply grafts a discrete rule onto the pre-existing legislative systems, with the assumption that social security decisions will involve the exercise of human cognition, whether expressed by a public servant, an officer acting on their behalf, or a computer program under their responsibility or control.

Those provisions also fail to provide any meaningful regulation of the use of AI: they simply assert that computerised processes can be used, and provide no further rules or principles regarding their use.

The GDPR

The major exception to the piecemeal approach to regulating AI technologies are data protection laws. In this section we emphasise the most prominent of those, the GDPR, which imposed a series of legal norms built around protecting "the fundamental rights

[2] *Social Security Act 1998* (UK).

[3] *Social Security (Administration Act) 1999* (Cth).

and freedoms of natural persons and in particular their right to the protection of personal data".[4] It is worth noting that the establishment of the right to data protection as a fundamental right is a distinguishing feature of European law.[5]

While the GDPR appears to have a broad scope, it still illustrates the piecemeal approach to the legal regulation of AI use by governments.

The GDPR's major provisions regulate the 'processing'[6] of 'personal data'[7], ensuring that this is done either with the consent of the person to whom the data relates or other legitimate authority, including in the case of government where data processing is 'necessary for the performance of a task carried out in the public interest or in the exercise of official authority'.[8] Higher thresholds apply to sensitive data (racial or ethnic origin, political opinions, religious beliefs, biometric data, data concerning health, sex life or sexual orientation), including the need to obtain 'explicit consent' from the person to who the data relates and for governments which seek to process information non-consensually to provide safeguards for the protection of data in legislation.[9]

A relatively wide spread of legally enforceable remedies are provided under the GDPR, including: rights to access personal data; erasure of personal data (right to be forgotten); and rectification (correction of inaccurate personal data).[10] While those legal institutions are valuable regulatory mechanisms, individually or together, they do not seek to regulate the use of AI beyond basic data protection safeguards.

The first limitation of the GDPR is implied by its name: ultimately it is a data protection framework and is primarily concerned with preventing private enterprises and governments from collecting and using citizens' data in non-consensual and harmful

[4] Art 1.

[5] This has been the case since the entry into force of the Lisbon Treaty in Dec 2009. See https://global.oup.com/academic/product/the-foundations-of-eu-data-protection-law-9780198718239?cc=au&lang=en&

[6] Art 4(2) 'collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction'

[7] Art 4(1) 'any information relating to an identified or identifiable natural person ('**data subject**'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person'

[8] Art 6(1)(e)).

[9] Art 9.

[10] Arts 16-21.

ways. In that sense, it does not attempt to provide comprehensive regulation of the enhanced powers which AI confer on governments.

The second limitation is that, even within the data protection sphere, the GDPR leaves enormous scope for governments to avoid a number of substantive legal protections. Governments may restrict data protection rights where necessary and proportionate to safeguard national security, defence, public security, the prevention, investigation and prosecution of crime, the protection of judicial process, and other objectives of general public interest including monetary, budgetary and taxation matters, public health and social security.[11] Even the most sensitive data can be processed without consent where processing is necessary in the field of employment and social security and social protection law, or in the public interest in the area of public health.

That long list of exemptions covers most of the major activities of governments, thereby creating the possibility for the wholesale disapplication of much of the GDPR's data protection rules to the public sector, provided that any restriction 'respects the essence of the fundamental rights and freedoms and is a necessary and proportionate measure in a democratic society to safeguard'. The precise meaning of that limitation on governments' powers to exempt themselves from the GDPR will vary significantly in different contexts.

Soft law preference

The third defining feature of the current legal landscape applying to the use of AI in the public sector is the use of 'soft', rather than 'hard', law to regulate AI.

Soft law comes in several types, prominently including:

1. Legal rules which impose no limitations or constraints and no significant remedial impact, i.e. no/low compensation and non-coercive court orders.

2. Non-legal rules (such as self-regulatory or industry-administered regimes) which rely on commercial or social incentives to enforce behaviour.

Both types of soft law have been normalised in the regulation of AI by government. Rather than produce robust legislation to address the power imbalance created by the use of AI by government, the principal regulatory response has been to adopt non-legal 'guides' and 'standards' of 'ethical' rather than 'legal' force.

Examples of that regulatory posture abound:

- The absence of heavy financial penalties for breaches of privacy laws which are proportionate to the economic resources of major technology firms and governments.

[11] Art 23.

- The Australian government's 'AI Ethics Framework' which provides merely voluntary principles for business and government (See: "Australia's Artificial Intelligence Ethics Framework").

- The Canadian government's 'Directive on Automated Decision-Making' which includes some requirements concerning transparency of the use of automation, but describes 'severe consequences' for breach as including 'Direct Cabinet discussion' (for agencies) and 'no performance pay' (for individuals) (See: "Directive on Automated Decision-Making").

- The UK government's 'A guide to using artificial intelligence in the public sector' published by the Government Digital Service and the Office for Artificial Intelligence which advises government officials to 'establish ethical building blocks for the responsible delivery of your AI project' but provides no indication of sanctions for breaching existing law (See: "A guide to using artificial intelligence in the public sector").

Toothless legal frameworks and the turn to 'AI ethics' illustrate the prevailing preference of regulators to leave major AI technologies outside the formal legal process.

That preference for dealing with government use of AI outside hard legal frameworks provokes obvious questions: if AI confers such extraordinary powers on governments, why is it not regulated by hard legal frameworks? What impact does the preference against hard legal governance have on the legitimacy of government uses of AI in political communities which aspire to liberal democratic values?

The next Part III sets some intellectual standards against which to begin addressing those queries. It provides a set of six criteria against which the legitimacy of legal frameworks governing public sector AI can be measured: the 'Audit Criteria' which are built from basic requirements of liberal democratic government. The following Part IV deploys those Audit Criteria in the context of four case studies in which legal controversies have arisen concerning government use of automation, machine learning, data archiving/networking and mass surveillance.

Part III: Audit Criteria

In this Part, we present the standards against which the current legal frameworks governing the use of AI in the public sector are assessed: the **Audit Criteria**.

First, we explain how (and why) we derive those standards from basic requirements of constitutional liberal democracy. We then state and explain the six Audit Criteria: **Knowledge**, **Assent**, **Personhood**, **Basic Protections**, **Contestability** and **Remedial Action**.

We close the Part by explaining how our Audit Criteria differ from existing methods for assessing the value of legal frameworks applying to AI:

1. Regulatory approaches aimed at boosting economic productivity and ensuring public safety;
2. Human rights approaches to AI; and
3. The sub-fields known as 'Fairness, Accountability and Transparency' and 'AI Ethics'.

Requirements of liberal democracy

Our Audit Criteria are drawn from a basic set of political and social requirements of constitutional government in liberal democracies.

Democracies are 'liberal' when they give overriding priority to the personal freedom of individual citizens: when they protect the 'liberty' of people to decide the rules which will govern them and a core set of personal rights (such as rights to privacy, free expression, assembly, conscience and property). Respecting those liberties does not entail a society free of responsibilities or solidarity, although it does require that government officials treat individual human beings as unique and valuable: as 'ends in themselves' rather than 'means to an end' (See: "Freedom in the World Research Methodology").

Those requirements of liberal democracy can be expressed as a set of practical institutional requirements, **Liberty Requirements**:

- ➢ Free elections in which citizens choose their representatives and, sometimes, vote on specific legislation ('consensual' government or 'self-rule') (See: "Monitoring Human Rights in the Context of Elections");

- Rules which keep government officials within the boundaries of parliamentary legislation (avoiding 'domination' of citizens by 'arbitrary' power);[12]
- Judicial mechanisms to enforce those rules which are independent of government officials ('rule of law');[13]
- Legal rules protecting a basic set of liberties, including:[14]
    - rights to vote in elections;
    - right to equal treatment/non-discrimination on the ground of race, gender, sexuality, political opinion, age and physical/intellectual ability;
    - rights to a private life (separate to a public life);
    - freedom to speak, assemble and intellectual opinion or 'conscience';
    - right to personal property, being economically valuable items which are self-generated; and
    - rights to be coerced (by government officials and private persons) only through valid parliamentary legislation; and
- Only derogating from the above requirements where the survival of the body politic is at stake, i.e. a very limited 'state of emergency'.[15]

While the Liberty Requirements may appear trite to academic commentators, they provide the **universal** and **overriding** standards for assessing whether a given legal (or social) rule is compatible with (or violates) the core values of modern societies.

- The Liberty Requirements are 'universal' because they must exist in all societies which wish to be called liberal democracies: whether or not they have constitutional human rights protections (A good example of the universality of these rules can be found in the Freedom House, Freedom in the World Research Methodology 2019).

[12] See eg, *Universal Declaration of Human Rights*, GA Res 217A (III), UN GAOR, UN Doc A/810 (10 December 1948) Art. 9, 12, 15, 17 ('*UDHR*').

[13] See, eg, *The rule of law and transitional justice in conflict and post-conflict societies: Report of the Secretary General*, UN Doc S/2004/616, [2]-[8]

[14] See the list in the *Freedom Report Methodology 2020*.

[15] Office of the High Commissioner for Human Rights, 'Chapter 16: The Administration of Justice During States of Emergency' in Human Rights in the Administration of Justice: A Manual on Human Rights for Judges, Prosecutors and Lawyers (United Nations Publications, 2003) p 813-816.

- The Liberty Requirements are ‘overriding’ because they sit above individual legal rules and provide a yardstick against which to measure the legitimacy and value of existing and proposed legal frameworks.[16]

In order to use the Liberty Requirements as the basis of this Legal Audit, we express them as six Audit Criteria:

1. **Knowledge**
2. **Assent**
3. **Personhood**
4. **Basic Protections**
5. **Contestability**
6. **Remedial Action**

[16] Although many of those requirements are protected through legal rules: for example, rights of privacy may be protected through legal rules which require warrants before conducting searches or entering property.

Criterion 1: Citizen knowledge of AI

The first Audit Criterion is the ‘**Knowledge Criterion**:

> ***citizens must possess sufficient knowledge about how a given AI technology uses information and achieves outcomes in order to assent (through the legislative process) to its use by government.***

Expressed as a question:

> ***does a person who is legally entitled to vote have sufficient knowledge about how technology is specifically used by government in order to decide whether to assent to its use.***

The Knowledge Criterion is the logical starting point for an assessment of the legitimacy of legal frameworks governing the use of AI by governments. Before a citizen can decide whether to assent to a particular use of AI, they must understand its fundamental technical basis and the specific impact it can have on their lives (For instance, consider; “Artificial Intelligence Technologies and Freedom of Expression; Factsheet 3”).

Without that knowledge, the fact that a person has shared information with government agencies or commercial actors does not justify it being used in particular way, and therefore will not provide any meaningful ‘self-rule’.[17]

Because the technologies which amount to AI are augmentations of human cognition, “sufficient knowledge” about those technologies should be measured on an equivalent basis to human decision-making:

> *does an 18-year-old intellectually-competent citizen understand how a particular technology uses information and achieves outcomes to the same level as they understand how a human decision-maker uses information and achieves outcomes?*

Answering that question does not require an ordinary person to have advanced knowledge about neuroscience because most humans have intuitively correct understandings of the basic reasoning processes of other humans:

> *‘if someone says I must be punished for breaking a rule, they must tell me what the rule is, show me evidence of how I broke it and explain the connection between the rule breaking and punishment’.*

In order to meaningfully assent to living in a society where the exercise of public power is assisted by the use of AI, ordinary people must be presented with a sufficient level of

[17] *Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression*, UN Doc A/73/348 [37]-[52].

knowledge to be able to explain how AI can be deployed in the same reasoning process.[18]

Meeting that standard does not require a degree in computer science, but it does require the translation of technically complex processes into plain English text.

In practical terms, the Knowledge Criterion places obligations on governments wishing to use AI technology in exercising public power to publicly disclose features such as: 1) what data has been used to inform the AI technology, 2) what the AI is trying to optimise, 3) how it reaches conclusions, 4) the outcome of the AI system, and 5) the effect on rights, in a way that is meaningful to a specific person given their specific situation and knowledge, and also functional for that person to be able to contest the outcome.

Criterion 2: Citizen assent to AI

The second Audit Criterion is the '**Assent Criterion**:

> ***citizens must assent to the specific use of AI before it is used by government.***

Expressed as a question:

> ***have citizens assented (through the legislative process) to the use of a particular type of AI technology being applied to a particular context?***

The Assent Criterion arises because the augmentation and diminution of human cognitive, physical and social capacities enabled through AI creates a new balance of power between public officials and ordinary citizens which must be considered and approved afresh by individual citizens in a liberal democracy.[19]

The ordinary way for people to assent to governmental action is through their elected representatives in parliaments voting for legislation (Consider; OECD, "Government at a Glance 2017*")*. Thus, the Assent Criterion requires specific legislative authorisation – beyond bare authorisation – before AI may be used by a government official.

Assent to the use of AI must be 'meaningful', rather than providing a simplistic or vague legal authorisation to use 'software' or 'computer systems'. That type of crude legal authority would neither specify the different types of AI technologies which can be used by government, nor recognise the different types of goods and harms which those technologies provide.

[18] Consider the Australian Government Commissioned CSIRO Ethics Report; D Dawson et al, 'Artificial Intelligence: Australia's Ethics Framework' (Data 61 CSIRO, 2019) Pt 3.1 ('*Australia's AI Ethics Framework*').

[19] *Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression*, UN Doc A/73/348 [37]-[52].

In order to provide that meaningful assent, legislation authorising the use of AI must:

- Specify the type of technology being used;
- Describe the technology and how it operates to use information and achieve outcomes by a standard capable of verification and contestation; and
- Declare its benefits and potential harms in more than gestural terms.

Each specification, description and declaration must be expressed in terms which an ordinary citizen could understand. Without that level of clarity, there can be no meaningful connection between the knowledge of AI possessed by ordinary people, and the assent given through their representatives via the legislative process.

Criterion 3: Personhood

The third Audit Criterion is the '**Personhood Criterion**':

> ***Governments must respect the diversity, autonomy and individual choices of persons whose lives are affected by the use of AI.***

Expressed as a question:

> ***Does a government's use of AI treat a person as a unique individual, with the capacity to make diverse and autonomous choices about their lives?***

The Personhood Criterion arises from requirements that governments must treat people within their societies as individual persons with autonomous decision-making capacities and free choice. It is strongly implied by core legal and constitutional principles, including:[20] constitutional principles of dignity, liberty and representative government; natural justice and due process requirements that each person affected by government action must be given an opportunity to contest that action; and non-discrimination norms that prohibit governments from taking adverse action against groups of individual persons who belong to similar cultural, biological or social groups.

The Personhood Criterion raises particularly acute issues in the context of government use of AI, because many AI technologies operate without explicit information about an individual person, relying on an assumption that individual people's behaviour will mirror the behaviours of historical groups of people with similar attributes to the targeted individual. As Part IV explains, automation and machine learning technologies are high-profile examples of those types of technology.

Positive satisfaction of the Personhood Criterion requires that AI systems used by public officials are designed and operate in a way which takes account of the unique characteristics, actions and behaviours of each individual person affected by their use.

Obvious violation of the Personhood Criterion would be evidenced by AI systems that adversely affect people's rights and interests by reference to inferences from historical data sets which assume that any given person has or will behave in an identical way to the behaviours of similar but distinct persons in the past.

[20] See, eg, Article 1 of the German Basic Law (*Grundgesetz für die Bundesrepublik Deutschland*), 14th Amendment of the *United States Constitution,* Article 1 of the French Constitution of 4 October 1958, Section 15 of the *Canadian Charter of Rights and Freedoms*; *Ridge v Baldwin* [1964] AC 40; *Plaintiff S157/2002 v Commonwealth* (2003) 211 CLR 476.

Criterion 4: Basic Protections

The fourth Audit Criterion is ‘**Basic Protections Criterion**’:

> ***the basic liberties of citizens must be protected by legal rules which apply to the use of AI by government.***

Expressed as a question:

> ***are the basic liberties of citizens protected by legal rules which apply to the use of AI by government?***

All countries wishing to be described as ‘liberal democracies’ must provide legal mechanisms to protect the basic political and civil rights of individual citizens,[21] most prominently including those expressed in the Liberty Standards with which we commenced this Part.[22]

Those rights must be explicitly protected through legal rules to meet the Basic Protections Criterion. Such legal protections can be provided by a single document (such as a Bill or Charter of Human Rights) or by separate legal rules dispersed throughout the legal system (such as laws limiting the powers of police to search premises, or judicially-created doctrines which protect free speech or personal liberty).[23]

However they are expressed, the legal protections of those basic rights must apply to uses of AI just as readily as they apply to human behaviour undertaken without technological augmentation. Ideally, those protections would be legally linked to the specific legislative authorisation of the use of AI.

[21] *Freedom in the World Research Methodology.*

[22] Eg *Further measures to promote and consolidate democracy*, Commission on Human Rights Resolution 2002/46; Art. *UDHR* Arts 2, 17, 18, 21, 27, 29.

[23] Eg George Williams, *The Federal Parliament and the Protection of Human Rights* (Department of the Parliamentary Library, 1999); Leslie Zines, ‘A Judicially Created Bill of Rights Symposium: Constitutional Rights for Australia’ (1994) 16(2) *Sydney Law Review* 166; Compare different approaches on each protection provided; Centro de Investigaciòn y Capacitaciòn Propuesta Civica A.C et al, How to Create and Maintain the Space for Civil Society: What Works? (‘‘How to Create and Maintain the Space for Civil Society: What Works?’’).

Criterion 5: Citizen contestability of AI

The fifth Audit Criterion is the '**Contestability Criterion**'

> ***citizens must be able to contest the legality of every use of AI by government.***

Expressed as a question:

> ***can citizens contest the use of AI by government through legally enforceable mechanisms?***

In order to ensure that governments use of AI falls within the scope of citizens' assent and their basic rights, it is necessary to provide an institutional mechanism which is independent of the government and capable of understanding the technological fundamentals of AI and its application.

At a minimum, the Contestability Criterion requires judges to adjudicate on the lawfulness of government's use of AI. Principles of the separation of powers doctrine ensure that judges are independent of public officials and able to protect citizens from unlawful government behaviour.[24] In that sense, the Contestability Criterion will usually be satisfied by rights to sue a government before an independent judge.

The existence of judicial review of government power does not, however, exhaust the requirements of the Contestability Criterion.

Full satisfaction of the Contestability Criterion requires the following institutional features:

- *Technical education/advisor*: judges must be educated in the basic features of AI technologies and their application. Without that education judges are not in a position to determine whether the (often opaque) operation of AI technologies fall inside/outside the scope of citizen assent. Education may take the form of an independent expert who is appointed to advise the judges on technical matters.
- *Speed*: people must have access to justice in a sufficiently time-effective manner because the speed and scalability of AI means that unlawful uses of AI can cause enormous harm unless quickly checked.
- *Accessibility*: procedures must exist in judicial enforcement bodies which permit the technology underlying AI to be accessible to both people challenging its lawfulness and technical experts retained by those people.

[24] Office of the High Commissioner for Human Rights, 'Chapter 14: Independence and Impartiality of Judges, Prosecutors and Lawyers' in *Human Rights in the Administration of Justice: A Manual on Human Rights for Judges, Prosecutors and Lawyers* (United Nations Publications, 2003) p 115-122.

Criterion 6: Remedial Action for wrongs committed by government use of AI

The sixth Audit Criterion is the '**Remedial Action Criterion**'

> ***citizens must have access to remedial action to correct and compensate for harm caused by the use of AI by government.***

Expressed as a question:

> ***are legal remedies available which provide compensation for harms caused by the use of AI by government?***

Liberal democratic government requires that illegitimate government actions be remedied by monetary or coercive orders (See: International Covenant on Civil and Political Rights 1966, Art. 2.3; Nicolaidis and Kleinfeld 2012, pp. 54-55):

- ➢ Monetary orders which compensate a person for harms suffered at the hands of public officials;
- ➢ Restitution of property wrongfully obtained or used by governments;
- ➢ Injunctions preventing continuation of illegal government behaviour; or
- ➢ Criminal penalties against public officials for seriously harmful behaviour.

Applied to AI, the Remedial Action standard requires those types of remedies to be applied to uses of AI by public officials.

In that context, meeting the Remedial Action Criterion may require legally enforceable orders of the following kinds:

- ➢ Orders for cleaning public records of data about a person which was generated through the wrongful use of AI;
- ➢ Orders for return of personal information which was wrongfully obtained; and
- ➢ Orders for compensation which are reflective of the economic, emotional and social impact of illegitimate use of technology.

Other methods of evaluating legal governance of AI

Auditing the law governing the use of AI by government using the Audit Criteria we have selected (Knowledge, Assent, Personhood, Basic Protections, Contestability and Remedial Action) differs from other methods of evaluating the legitimacy or desirability of legal regulations of AI.

A variety of different approaches have been adopted by regulators and commentators for measuring the quality of laws to govern AI, including focusing on:

- Economic productivity or public safety;
- ‘Fairness, accountability and transparency’;
- ‘AI ethics’; and
- ‘Self-regulation’.

Before moving to the substance of the Legal Audit, we explain our reasons for choosing not to adopt those approaches.

Economic productivity/public safety

A popular way of assessing the success of the legal regulation of AI is to focus on the potential impacts of that regulation on:

- Economic productivity (See: “Industrial Strategy: Artificial intelligence Sector Deal”; or
- Public safety (See: “AI: Using Standards to Mitigate Risk”).

Assessing legal regimes governing AI in that way is meaningfully different to other approaches which assume that legal regimes governing AI will invariably trade-off economic development or public safety against other societal aims, such as representative government and the protection of liberties.

Our inquiry is a necessary *precondition* to the productivity/public safety approaches to law and AI. In general, the basic principles of liberal democracy do not assume any final balance of personal liberty and countervailing social goals, such as economic development and public safety. Parliamentary legislation creates the basic institutions for economic activity (e.g., central banks, courts to enforce contracts) and public safety (e.g., police and military forces). Through those legislative institutions, ordinary citizens can choose to trade-in some of their liberty for a safer or richer society.

But the basic principles of liberal democracy do require that a particular process is followed in deciding whether money or safety should override certain basic democratic liberties.

Specifically, reasonable decisions to adopt a particular trade-off between personal/social liberties for economic benefits/public order involve (explicit or implicit) engagement with a number of topics:

- How and why liberty could be reduced by adopting AI in order to realise economic gains or increase social safety (Knowledge Criterion);
- Whether liberty should be reduced in that way (Assent and Personhood Criterion);
- Whether the terms of their assent have been breached (Contestability Criterion) and remedying the breach in a proportionate way (Compensation Criterion).

Using our Audit Criteria, the question is not “does law X impose unfairly large burdens on developers of information technology in order to safeguard against a potential infringement of liberty”. Rather, the question is “does the legal regime applying to these technologies provide an opportunity for people to express a liberal democratic view on whether commercial objectives should override liberty?”

Fairness, accountability and transparency

Another influential academic approach to the regulation of AI technologies has been developed in academic literature which focuses on the ‘fairness, accountability and transparency of machine learning’ or **FATML**.

The FATML approach to the regulation of AI is to provide debate, rules and standards to guide the development of ethical algorithmic systems (including: "Principles for Accountable Algorithms"):

> ***Responsibility***: *Make available externally visible avenues of redress for adverse individual or societal effects of an algorithmic decision system, and designate an internal role for the person who is responsible for the timely remedy of such issues.*
>
> ***Explainability***: *Ensure that algorithmic decisions as well as any data driving those decisions can be explained to end-users and other stakeholders in non-technical terms.*
>
> ***Accuracy***: *Identify, log, and articulate sources of error and uncertainty throughout the algorithm and its data sources so that expected and worst case implications can be understood and inform mitigation procedures.*
>
> ***Auditability***: *Enable interested third parties to probe, understand, and review the behavior of the algorithm through disclosure of information that enables monitoring, checking, or criticism, including through provision of detailed documentation, technically suitable APIs, and permissive terms of use.*

> ***Fairness****: Ensure that algorithmic decisions do not create discriminatory or unjust impacts when comparing across different demographics (e.g. race, sex, etc).*

While the FATML approach to the regulation of AI is valuable, it differs from our inquiry in important ways. The FATML movement is not primarily directed towards critique or reform of existing legal frameworks. Instead it has been intensely focused on questions of computational fairness, with a minor emphasis on transparency, and virtually no attention to accountability. The FATML principles speak mainly to software and system engineers, not to governments and their power dynamics citizens.

'AI ethics'

Another common way of approaching the regulation of AI technologies is by focusing on 'ethics' and articulating special ethical rules described as 'AI Ethics'. Originating in academic circles, AI ethics has become a prominent part of regulatory debates surrounding AI generally, and AI used by governments (For a summary of the academic position, see: "Ethics of Artificial Intelligence and Robotics").

In some cases, government agencies have promulgated ethical principles for the use of AI (See: "Australia's AI Ethics Framework"; CSIRO 2018), or adopted academic ethical debates as the prism through which legal institutions which govern AI should be conceived (See: "Human Rights and Technology: Discussion Paper"). Those governmental adoptions of 'ethics' as an appropriate lens to view the regulatory challenges of AI follow the apparent preference of large technology companies to move debates around the legality of AI towards debates around 'ethics' (See: Microsoft, "Responsible AI"; KPMG 2019).

We do not consider 'ethics' to be the most appropriate way to think about the legal regulation of government use of AI, although we do agree that the engagement of AI with ethical norms produces valuable and important academic and popular debates.

Modern governments obtain legitimacy through providing concrete and democratically accountable avenues to challenge harmful behaviour: principally, via legal institutions such as parliamentary legislation and independent judicial bodies. While ethical considerations surely feed into the decision-making of parliamentarians and judges, concentrating on 'ethics' rather than 'law' in the regulation of AI in government is an unhelpful distraction. Fixating on 'AI Ethics' diverts attention from concrete, practical, solutions to harmful uses of technology, towards abstract academic topics which are designed to be endlessly debated and are unlikely to produce timely options for institutional reform.

Part IV: Case Studies

This Part provides an audit of various legal frameworks governing the use of AI by government agencies and public officials. It spans several jurisdictions within the OECD and focuses on one case study for each type of AI technology:

- ➢ **Automation**: the automated debt collection technology used by the Australian government (**robo-debt/OCI**);
- ➢ **Machine learning**: the criminal risk prediction machine learning system used throughout the US criminal justice system (**COMPAS**);
- ➢ **Data archiving/networking**: the bulk transferral of archived patient data by the UK's national health service to a Google subsidiary (**DeepMind**); and
- ➢ **Mass surveillance**: the live automated face recognition technology used by the UK police force (**NeoFace Watch**).

Those case studies are selected for the following three reasons:

1. **Direct legal challenge**: the legality of the relevant use of
2. AI technology was directly challenged, permitting the legal rules governing AI to be audited 'in context' rather than 'in the abstract';**Technological transparency**: the basic technical features of the AI technology were revealed, either through dispute resolution processes or journalistic intervention; and
3. **Prominent impacts**: the human impacts of both technology and law were revealed.

The analysis of each case study follows the same sequence. The basic facts are presented, focusing on the type of AI technology used, the legal rules which applied to the use of that technology, how those legal rules operated, and what happened to ordinary citizens.

After that analysis, each of the six Audit Criteria are applied:

1. Knowledge
2. Assent
3. Personhood
4. Basic protections
5. Contestability

6. Remedial Action

A numerical score is given after applying each of the six Audit Criteria.

(0) No compliance	(1) Weak compliance	(2) Moderate compliance	(3) Strong compliance	(4) Excellent compliance

Automation

The first type of law to be audited applies to the use of automation technologies by governments.

As Part I explained, automation technology is most easily understood as a type of advanced calculator which works by applying logical rules to inputs supplied by a user in order to produce outputs:

if input = X → produce output Y, otherwise → produce output Z.

The case study selected to audit the compliance of legal regimes governing automation technologies is drawn from the use of wholly-automated debt-collection procedures adopted by the Australian government's social welfare agency.

Robo-debt collection

In 2015, the Australia government adopted a fully-automated system for recovering amounts of money (which it identified as 'debts') from recipients of social welfare: the 'Online Compliance Intervention System' (**OCI**) (The following account of the design and operation of OCI is taken from 2 reports published the Commonwealth Ombudsman in 2017 and 2019: "Centrelink's automated debt raising and recovery system" (2017); "Inquiry into Centrelink's compliance program" (2019)).

Data-matching technology had been used in the Australian social welfare agencies since 2011: taking employer information about employment income and comparing it to information voluntarily disclosed by a welfare recipient on a rolling monthly/quarterly basis. That system identified potential overpayments through "***averaging***" monthly/ quarterly income into estimated weekly income. That automated averaging technology was initially used as a mechanism to alert human public officials to potential fraud: leaving the decision about whether to impose a penalty on recipients of welfare payments to legally responsible humans. From 2015, the humans were removed from the process and all debt recovery actions were automated via the OCI system, the technical details of which are explained below. The result was a stunning increase in the number of penalty notices: moving from 20 000 a year to 20 000 a week.

Between 2016-2019, Australian administrative tribunals warned that the use of the OCI system was obviously illegal, on the basis that enforcing a debt solely based on an individuals' predicted income (via the automated program), rather than evidence of actual payments received by a person, fell outside the terms of the relevant legislation (The tribunal decisions are embedded in: "Coalition warned robodebt scheme was unenforceable three years before it acted").

Despite those non-judicial rulings, under the OCI system (See: "Robodebt: government to refund 470,000 unlawful Centrelink debts worth $721m"):

- unlawful demand letters were sent to 373,000 people;
- $A721million was unlawfully demanded by (and paid to) the Australian government as a result of the use of OCI.

In 2018, an Australian legal aid agency brought a public interest lawsuit arguing that the use of OCI was illegal. In late 2019, the Federal Court of Australia ruled that OCI was unlawful on the ground that:

- the relevant human decision-maker with legal responsibility for debt collections 'could not have been satisfied that a debt was owed in the amount of the alleged debt' because the OCI produced only averaged, rather than actual, data about citizens' income.
- a 10% penalty could not be added to the debt unless an public official had formed a view regarding the independent merits of the case.

Since that ruling, the Australian government has indicated it will repay the funds collected through OCI. A civil class action has been launched seeking compensation for all people who had money unlawfully collected via OCI. There is currently no public decision on whether the use of OCI will be discontinued, although the Australian social welfare agencies have indicated that they have tweaked the parameters of the software

Technology

OCI was a clear example of 'automation' as a type of AI technology. It had the following technical features:

1. Automatic collection of citizen data from Australian taxation office concerning income;
2. Arithmetic calculation by averaging annual income across monthly/quarterly periods;
3. Identification of an overpayment by reference to the averaged periods; and
4. Automatic issue of demand letters and eventual debt recovery.

Prior to the adoption of OCI, each of those steps was performed by a human employee of the Australian welfare department, who adopted the following process before demanding a citizen pay money to the government:

1. income data collected from the taxation office was used to identify whether a person had been overpaid their social welfare entitlement;
2. a human public official would communicate with the person seeking clarification concerning the potential overpayment (by way of proof of income, such as payslips from their employer and other physical evidence):

    a. If a response was received, the human official entered the actual amounts paid into a calculation system to determine if there had been an overpayment;

        i. If there was no overpayment, the matter ended there; and

        ii. If was an overpayment, the welfare recipient was sent a letter requiring repayment of the debt.

3. if no response was received, the human official would write to the person's employer (or relevant third party) seeking confirmation of the exact amounts paid to the person over the relevant timeframe :

    a. if a response was received from the employer, the human official entered the actual amounts paid into a calculation system to determine if there had been an overpayment; and

    b. if no response was received from the employer, the human official could apply 'averaging' software to the income information from the taxation office to estimate whether there had been an overpayment.

The differences between the automated and human processes were stark. The automated OCI system:

- ➢ relied entirely on estimation rather than actual information regarding a person's compliance with the welfare legislation;
- ➢ never attempted to verify whether financial data provided by an employer was accurate; and
- ➢ never included a politically or institutionally response human in the process of demand repayment of a welfare benefit.

Impact on ordinary people

OCI's impact on ordinary people was enormous.

While OCI surely delivered significant cost savings and additional cashflows to the Australia government, its use caused significant harm to the people who received automated demands for payment. Those harms included:

- ➢ imposes unlawful financial burdens on vulnerable people;
- ➢ increasing emotional burdens on those people;
- ➢ creating illegitimate social stigma; and
- ➢ creating feelings of powerlessness in ordinary people in their dealings with government agencies.

Ultimately, OCI imposed heavy financial and emotional burdens on welfare recipients without any legal justification, as well as a feeling of powerlessness in the people who received unlawful demands for payment. Many of the people who received unlawful demands from the OCI program fell into vulnerable cohorts.

The use of OCI also diminished trust in the Australian government. +300,000 (representing +1% of the Australian population) unlawful demand letters issued by the OCI system.

Legal rules

OCI was deployed in a legal context which contained a mixture of legislative and judge-made law.

The legislative framework governing social security in Australia was (and is) complex, but the core provisions relevant to the legality of OCI can be distilled:

- Welfare payments would only be made to people who earned under a certain amount of fortnightly income.
- If more income was earned, the person was no longer entitled to welfare.
- Any welfare paid in excess of the entitlement became a ‘debt’ due to the government.
- An additional financial penalty could be imposed if a person failed to provide information regarding their income.
- The government official administering the welfare payment could decide to waive the financial penalty if the person had reasonable cause for failing to provide relevant information.
- A senior public employee could ‘authorise’ the use of a ‘computer’ to make decisions, although there was no requirement to publish the authorisation or specifically authorise the processes used by the authorised computer.

More general principles of public law also applied to the legality of OCI:

- Liability to re-pay welfare benefits (ie, a ‘debt’) only arose if a welfare recipient had actually received income in excess of the relevant threshold.
- The decision to waive a penalty had to take account of each welfare recipient’s individual circumstances, rather than invariably applying a general rule (the “**no fettering rule**”).
- A person who wished to contest their ‘debt’ to the government could challenge the decision to collect the debt and the financial penalty in a quasi-judicial body: the “**administrative appeals tribunal**” or “**AAT**”.

- A decision of the AAT could be appealed to a fully-judicial body: the "**Federal Court of Australia**".
- A non-judicial complaints handling body could investigate the OCI system and decide whether there had been improper (but not necessarily, illegal) conduct: the "**Commonwealth Ombudsman**".

Law in operation

The Australian government's use of algorithmic debt-recovery was challenged or reviewed through four different mechanisms:

- Review 1: merits review before the Administrative Appeal Tribunal (2017) -> finding that OCI was unlawful: not finally legally-binding.
- Review 2: judicial review before the Federal Court of Australia (2018) -> ruling that OCI was unlawful: legally-binding.
- Review 3: class-action for monetary damages in the Federal Court of Australia (ongoing): Australian government refusing to pay full-compensation (interest and damages for distress).
- Review 4: the Commonwealth Ombudsman conducted an inquiry into the operation and propriety of the OCI system.

No review provided complete legal relief to ordinary people who received unlawful demands for payment via automation technology.

Review 1 was provided by a quasi-judicial tribunal, the Administrative Appeals Tribunal, which does not have power to order the Australian government to cease using an unlawful algorithm, but is limited to deliberating on the merits of individual cases where unlawful algorithms may have been used. The Tribunal found that OCI was unlawful (in individual cases) because the Australia welfare agency had no legal authority to demand repayment of money unless it had actual proof that a welfare recipient had obtained income in excess of the amount permitted by legislation. The use of averaging by OCI did not provide the agency with that actual proof, leaving the agency without any legal authority to demand re-payment of allegedly overpaid amounts, nor authority to impose a penalty. The Australian government complied with the tribunal's orders in individual cases, but declined (as it was legally-entitled to do) to cease using OCI and continued to collect money through unlawful algorithmic demands.

Review 2 was provided by a fully-judicial body, the Federal Court of Australia, with power to order that an individual use of OCI was unlawful. Over 4 years after it began using OCI, the Australian government conceded in this forum that the technology was unlawful, but did not entirely cease using the technology. The precise basis upon which OCI was conceded to be unlawful has not been disclosed to the public.

Review 3 was also provided by the Federal Court of Australia, in its jurisdiction to order the Australian government to provide compensation to people suffered loss through the unlawful use of automation technology. The Australian government continues to contest its liability to pay full compensation for harm caused by OCI, over 5 years after it began using it.

Review 4 was provided by the Commonwealth Ombudsman: a non-judicial body with no coercive authority which investigates instances of potentially inappropriate government behaviour and makes recommendations to make government activities fairer. The Commonwealth Ombudsman reported on the use of OCI in 2 major reports, which found that the design of OCI was unfair in significant ways (See: "Centrelink's automated debt raising and recovery system" (2017); "Inquiry into Centrelink's compliance program" (2019)). Only some of the Ombudsman's reform recommendations were adopted by the Australian government.

In July 2020, the Australian government voluntarily offered to repay the unlawfully demanded amounts ($A721million), but refused to make full compensation to affected citizens: interest on those amounts and to pay compensation for harm suffered (including opportunity costs and emotional distress) as a result of the unlawful use of automation technology. Given the 5 year gap (2015-2020) between the unlawful demands and the voluntary repayment, full compensation would significantly exceed the $A721million unlawfully paid.

Audit Criteria

Knowledge: no compliance (0/4)

The law applying to the use of OCI did not comply with the Knowledge Criterion.

There was no knowledge (outside the Australian government and its associates) that the process of debt recovery would be wholly automated via the OCI algorithm. Nor, obviously, was there knowledge of the internal metrics of the OCI algorithm and the unlawful demand letters produced by the OCI algorithm failed to properly notify the recipients that no (human) public official had reviewed their case and decided to proceed with enforcement action.

Assent: no compliance (0/4)

The law applying to the use of OCI did not comply with the Assent Criterion.

No parliamentary legislation authorised the use of the OCI algorithm, nor has any been introduced to provide such authorisation since the exposure of the unlawful basis of the automated AI technology.

The Australian Ombudsman has noted that the use of OCI was 'authorised' by a senior public official in the Australian welfare agency under legislation (See: "Centrelink's automated debt raising and recovery system" (2017) at [2.35]), but such an authorisation did not evidence compliance with the Assent Criterion for two reasons. First, the relevant legislation entirely delegated the decision to 'authorise' the use of OCI to a non-elected official, leaving no meaningful connection between democratic accountability and the use of automation technology. Secondly, the relevant authorisation was not publicly released, leaving no possible basis to challenge the use of the technology in a political or legal forum before its deployment in an unlawful manner.

Personhood: strong compliance (3/4)

The law applying to the use of OCI moderately complied with the Personhood Criterion.

The transition to an automated system cut out multiple levels of human engagement that would have ensured that the diversity, autonomy and individual choices of persons were respected. Instead, crude principles were applied in a fashion divorced from individual circumstance, compounding the consequences for vulnerable populations.

Those functions of the OCI system underlay its illegality, as Australian administrative law required the relevant welfare agency to determine whether a person had actually been paid above the relevant threshold, rather than relying on automated predictions of their behaviour.

Basic protections: weak compliance (1/4)

The law applying to the use of OCI only provided weak compliance with the Basic Protections Criterion.

The most basic right which was infringed by the use of OCI was the human right to private property: the unlawful demands for repayment of welfare benefits deprived recipient of their right to their monetary property.

Under Australian law, there was no requirement to provide full compensation until ordered to do so by a full-judicial body in expensive and high-stakes litigation. Despite the commencement of such litigation, no order for full compensation for breach of the basic right to private property has been made.

In those circumstances, there has been only weak compliance with the Basic Protections Criterion.

Contestability: Moderate Compliance (2/4)

The law applying to the use of OCI only provided moderate compliance with the Contestability Criterion.

From that chronology, it is clear that the legal frameworks for challenging the use of automation technology in Australia provided a (only) moderately effective avenue to contest the legality of the OCI algorithm.

Strong compliance with the Contestability Criterion would have required a legal framework which facilitated a swift and categorical end to the obviously illegal use of OCI: i.e., to have ordered the government to cease using the algorithmic system within several months of its commencement.

Compensation: Moderate Compliance (2/4)

The law applying to the use of OCI only provided moderate compliance with the Compensation Criterion.

After consistently losing litigation for almost 4 years, the Australian government has agreed to re-pay the amounts unlawfully demanded via the OCI algorithm. However, the Australian government continues to refuse to provide full compensation for the harm suffered by people who received unlawful algorithmic demands for money. Critically, no judicial order has yet required the Australian government to make such full compensation.

In that sense, the law governing the compensation for the unlawful use of automation technology only provides moderate compliance with the Compensation Criterion.

Strong compliance would require a legal-enforceable right to full compensation from the first time the unlawful use of OCI was detected.

<u>Total Score: 8/24</u>

Audit Criterion	(0) No compliance	(1) Weak compliance	(2) Moderate compliance	(3) Strong compliance	(4) Excellent compliance
Citizen Knowledge	X				
Assent	X				
Personhood				X	
Basic Protections		X			
Contestability			X		
Remedial Action			X		

Comment on comparable systems

It is likely that the low score given to the Australian law governing the use of automation technology in welfare debt recovery would be largely replicated in comparable jurisdictions.

The jurisdiction with the strongest regulation of automated government processes is the GDPR. Article 22 of the GDPR relevantly provides:

> *Article 22 – Automated individual decision-making, including profiling*

> *(1) The data subject shall have the right not to be subject to a decision based solely on automated processing, including profi*
>
> *ling, which produces legal effects concerning him or her or similarly significantly affects him or her.(2) Paragraph 1 shall not apply if the decision:*
>
> *…*
>
> *(b) is authorised by Union or Member State law to which the controller is subject and which also lays down suitable measures to safeguard the data subject's rights and freedoms and legitimate interests; or*

Article 22(1) would likely apply to the use of OCI because the decision to issue demand letters was 'solely based on automated processing'. However, it is possible that the exception provided in Art 22(2)(b) would have applied because: (i) the use of OCI was 'authorised' by a decision made under the Australian social security legislation; and (ii) that legislation provided an avenue to challenge the use of OCI before the relevant quasi-judicial tribunal (described as 'Forum 1' above).

The 'no compliance' score on the Knowledge and Assent Criteria is likely to be repeated in the US, EU, UK, Canadian and New Zealand given the absence in each of those legal systems of any requirement for (i) publication of the design and function of automation algorithms before use and (ii) specific legislative authorisation of the use of those algorithms.

The low compliance for the Basic Protections Criterion may not be replicated in the US, EU, UK, Canada and New Zealand: each of those jurisdictions has explicit human rights law which applies to government behaviour which uses automation. However, the degree of protection of the right private property may not be strong. The human right which was most obviously challenged by OCI was the right to private property, but in each jurisdiction those rights are subject to expensive and protracted litigation processes. When those slow and costly processes are compared to the likely speed and scale of deprivations of property (and other basic) rights, it is unlikely that the mere existence of human rights instruments provides strong compliance with the Basic Protections Criterion.

For the same reason, the moderate compliances scores for the Contestability and Compensation Criteria are also likely to be replicated in comparable jurisdictions.

Machine learning

The next set of legal rules are those applying to the use of ML technologies in the public sector.

As Part I explained, machine learning technologies use algorithmic processes to analyse large quantities of information (expressed in data sets) in order to classify and recognise patterns in historical data, and then to use those patterns to make probabilistic predictions about future actions. Essentially, those technologies use large historical data sets to arrive at educated guesses about future humans behaviour.

The case-study selected to audit the compliance of legal regimes governing machine learning is drawn from the use of algorithms to predict the likelihood of criminal offending (specially, recidivism) in the US, with a particular focus on the use of a machine learning algorithm in criminal sentencing.

COMPAS

From the early 2000s, a number of States in the USA began using a commercial software program called the Correctional Offender Management Profiling for Alternative Sanctions (**COMPAS**).

Created by a private company, Northpointe Inc now Equivant Inc, COMPAS is a risk assessment instrument that uses machine-learning algorithms to assess recidivism rates. It can be used before (in bail hearings or sentencing a criminal offender) and during (by a parole authority) the incarceration of a convicted criminal.

COMPAS came to prominence in 2013 during the sentencing of Eric Loomis for driving a car that had been used in recent shooting, and subsequently charged with attempting to flee an officer and operating a vehicle without owner's consent. Although none of those crimes carried mandatory prison sentences, Mr Loomis was punished by 6 years in prison, and 5 years extended supervision. COMPAS was used by the judge to assess the risk that Loomis would re-offend if he were not sentenced to imprisonment. COMPAS indicated that Mr Loomis was highly-likely to re-offend and the sentencing judge relied on that indication to imprison Mr Loomis.

Mr Loomis challenged the legality of his sentence, arguing that COMPAS violated his due process rights because its technical specifications considered trade secrete which were not disclosed to his lawyers (See: State v Loomis [2016] WI 68). He further argued that the use of COMPAS violated US due process rights because it discriminated against people based on their race and gender. The Wisconsin Supreme Court rejected those challenges, finding that 'consideration of a COMPAS risk assessment at sentencing does not violate a defendant's right to due process'. The Supreme Court of the United States refused to hear an appeal from that holding, effectively endorsing the use of COMPAS under the *US Constitution.*

COMPAS continues to be used widely throughout the US.

Technology

Although the precise details of the COMPAS algorithm are confidential to its commercial owners, it appears to be a machine learning system.

Upon arrest or sentence, around 137 datapoints are collected on an accused or convicted person by interview or automated filling from court records. The questions asked of the person include (See: COMPAS Sample Risk Assessment):

- "Was one of your parents ever sent to jail or prison?"
- "How many of your friends/acquaintances are taking drugs illegally?"
- "How often did you get in fights while at school?"
- "Do you agree/disagree:
  - "A hungry person has a right to steal"
  - "If people make me angry or lose my temper, I can be dangerous."

The COMPAS algorithm parses a data set of previous offenders to determine correlations between the accused/convicted person's answers and historical answers of previously accused/convicted people who did/did not re-offend. On the basis of the degree of correlation between the present and historical data, the COMPAS algorithm produces a risk score for the accused/convicted person which is then used by judges/probation officers/police to determine whether to imprison the person.

A study of COMPAS's accuracy (by its commercial owner) assessed its recidivism-risk scores as around 68% accurate: 18% better than a coin-toss. Independent researchers later assessed the accuracy of the COMPAS algorithm and found that it was far less accurate when the race of the accused/convicted person was taken into account (See: "Machine Bias"):

COMPAS -> Reality	White	African-American
Labelled Higher Risk -> Didn't Re-Offend	23.5%	44.9%
Labelled Lower Risk -> Did Re-Offend	47.7%	28.0%

Critically, none of the present or historical data used by COMPAS to produce the risk-score concerns race. Instead, the gap between the risk-score and the reality of recidivism appears to result from weightings applied to data points in the algorithm which correlate with race, but not necessarily with recidivism. The commercial owners of COMPAS have asserted that the algorithm is not racist (See: Dieterich et al. 2016; Equivant 2018), but also refuse to release the algorithm's technical specifications which prevents independent researchers from determining the precise reason for the racially-differentiated prediction errors (See: Wadsworth et al. 2018; Dressel and Farid 2018).

Impact on ordinary people

The use of COMPAS has potential beneficial impacts on ordinary people. If COMPAS accurately predicted high-recidivism, then members of society are protected from re-offending behaviour which adversely affects their physical, mental and economic interests. If COMPAS accurately predicted low-recidivism, then low-risk offenders can be released into society to the benefit of their close community (family, loved ones, friends, employers, employees) and relieve society of the economic burden of unnecessary incarceration.

There are, however, also very significant negative impacts of COMPAS.

The first set of negative impacts arises from the problems identified in the accuracy of COMPAS. If a high-recidivism risk score is biased against members of certain social and biological, then members of the broader society yield no meaningful benefits from their incarceration (there are no benefits in imprisoning a person who will not re-offend), and members of their close community (including the offender) are deprived of the benefit of the offender's presence in their lives. The obverse is true for biased low-recidivism risk scores, which expose members of society to physical, emotional and economic harm by failing to imprison people who will re-offend.

The second set of negative impacts arises irrespective of the accuracy/bias of COMPAS's recidivism risk-scores. Each time COMPAS is used to make a final decision on incarceration, a public official is deciding a person's liberty based on the past behaviour of similar, but not identical, people. Thereby, the individual autonomy and individual identify of each person to whom COMPAS assigns a risk score is de-valued. That de-valuation of individual autonomy has a number of negative impacts, including loss of trust in the legal system and unwarranted social stigma/praise as offenders (and their close communities) observe that the state is less concerned with their actual behaviour than with the past behaviour of similar (but not identical) individuals.

Legal rules

Unlike the use of OCI in Australia (Case Study I), the use of COMPAS in Mr Loomis' sentencing was unsupported by any legislation: ie, there was no legislation which expressly approving the sentencing judge's use of COMPAS to assess Mr Loomis's risk of recidivism.

For that reason, the legality of the use of COMPAS depended on US and Wisconsin constitutional due process rights, particularly the following legally-enforced rights:

- Right to be sentenced on an individual basis, rather than according to membership of a particular social group which COMPAS assessed as more likely to re-offend.[25]
- Right to be sentenced on accurate information, rather than potentially faulty information in the data-set underlying COMPAS.[26]
- Right to be free from discrimination according to gender/sex and race.[27]

Law in operation

The principal forum to challenge the legality of COMPAS was the trial and appellate judicial system provided by the State of Wisconsin and the US Constitution.

After pleading guilty to being the driver in a drive-by shooting, Mr Loomis was administered a COMPAS test and assigned a score indicating a high risk of recidivism. The sentencing judge placed heavy weight on that score in sentencing Mr Loomis to 6 years in prison, and 5 years extended supervision:

> *You're identified, through the COMPAS assessment, as an individual who is at high risk to the community.*
>
> *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 utilized, suggest that you're extremely high risk to re-offend.*

Mr Loomis lodged an objection to the sentencing judge's reliance on COMPAS, arguing that the judge had failed to consider *his conduct*, and had effectively punished him for the past conduct of other people which the COMPAS algorithm identified as similar in type of Mr Loomis. The sentencing judge refused the objection, asserting that the COMPAS risk assessment algorithm merely corroborated other factors in Mr Loomis' case, such as the violence of the offence, and the same sentence would have been imposed regardless of COMPAS.[28]

Mr Loomis then appealed to the Wisconsin Supreme Court on three grounds. First, that the use of COMPAS violated his due process rights to be sentenced on the basis of accurate information because there was no meaningful information before the sentencing judge regarding the operation of the COMPAS algorithm and the technical

[25] *Craig v. Boren*, 429 U.S. 190, 208-210 (1976); *State v. Gallion,* 270 Wis. 2d 535 (2004).

[26] *Gardner v Florida*, 430 US 349 (1977); *State v Skaff*, 152 Wid. 2d 48 (Ct. App. 1989).

[27] *State v. Harris*, 326 Wis. 2d 685 (2010).

[28] *Loomis* at [28].

specifications of COMPAS were not disclosed to Mr Loomis. Secondly, by relying on COMPAS the sentencing judge violated Mr Loomis' right to individual justice, because COMPAS did not arrive at a personal risk score, but a risk score of a cohort of people which included Mr Loomis. Thirdly, Mr Loomis contended that the use of COMPAS in sentencing unlawfully discriminated against him on the basis of his sex, because the algorithm assigned a higher risk of recidivism to men than women. A critical part of each individual ground of challenge was the refusal of the commercial owners of COMPAS to release the technical specifications of the COMPAS algorithm.

The Supreme Court ruled against Mr Loomis on each ground.

Accurate information in sentencing

The Supreme Court accepted that there was no meaningful information explaining to Mr Loomis how COMPAS assigned him a high-risk score, but was content to rely on a vague explanation of COMPAS's operation offered in promotional material produced by the commercial owner of COMPAS:

> *[54] 'Loomis is correct that the risk scores do not explain how the COMPAS program uses information to calculate the risk scores. However, Northpointe's 2015 Practitioner's Guide to COMPAS explains that the risk scores are based largely on static information (criminal history), with limited use of some dynamic variables (i.e. criminal associates, substance abuse)'*

Ultimately, the Court ruled the sentencing judge and Mr Loomis had the opportunity to understand how COMPAS assigned a high-risk score on the following basis:

> *[53] 'Although Loomis cannot review and challenge how the COMPAS algorithm calculates risk, he can at least review and challenge the resulting risk scores set forth in the report attached to the PSI.'*

> *[56] 'The circuit court and Loomis had access to the same copy of the risk assessment. Loomis had an opportunity to challenge his risk scores by arguing that other factors or information demonstrate their inaccuracy.*

Individualised sentencing

The Supreme Court also rejected Mr Loomis's argument that the use of COMPAS led the sentencing judge to fail to provide "an individualised sentence", but instead punished him for his membership of a statistical cohort which had an historical trend of recidivism.

Curiously, the Court accepted as accurate the following statement provided by COMPAS's commercial owner that:

> *[69]"[r]isk assessment is about predicting group behavior . . . it is not about prediction at the individual level … [a]n offender who is young, unemployed, has an early age-at-first-arrest and a history of supervision failure, will score medium*

> *or high on the Violence Risk Scale even though the offender never had a violent offense."*

Despite recognising the inherent difficulty in using COMPAS to determine an individual's (rather than a cohort's) recidivism risk, the Court reasoned that the sentencing judge's reliance on COMPAS was lawful because it was "helpful" in arriving at an individual sentencing:

> *[72] 'we disagree with Loomis because consideration of a COMPAS risk assessment at sentencing along with other supporting factors is helpful in providing the sentencing court with as much information as possible in order to arrive at an individualized sentence.'*

Ultimately, the Court imposed the following 2 "limitation" on the future use of COMPAS in criminal sentencing:

> *[98] "…risk scores may not be used: (1) to determine whether an offender is incarcerated; or (2) to determine the severity of the sentence."*

The key word is "determine" which, in that particular legal context, means that a sentencing judge can still use COMPAS to "decide" whether to incarcerate an offender and to "impose" a severe sentence, so long as the totality of the sentencing process is not delegated to the COMPAS algorithm.

Sex discrimination

The Supreme Court rejected Mr Loomis's argument that the use of COMPAS unlawfully discriminated against him on the basis of sex (male) on the basis that there was no evidence that the sentencing judge explicitly relied on biological sex as an independent factor in sentencing Mr Loomis.

Again, the Court recognised the powerful legal and factual problems with using risk assessment algorithms like COMPAS in sentencing. Factually, the Court acknowledged that COMPAS would likely assign Mr Loomis a high-risk score because he was male: [78] "there is statistical evidence that men, on average, have higher recidivism and violent crime rates compared to women". Legally, the Court recognised US Supreme Court precedent which held that penalising men at a higher rate than women would be unlawfully discriminatory: [79] "the principles embodied in the Equal Protection Clause are not to be rendered inapplicable by statistically measured but loose-fitting generalities concerning the … tendencies of aggregate groups."[29]

Despite recognising those problems, the Court rejected Mr Loomis's challenge on two bases. First, the Court contended that Mr Loomis had failed to prove (rather than argue) that the sentencing judge relied on his biological sex as a factor in increasing his sentence:

[29] *Craig v. Boren*, 429 US 190, 208-210 (1976).

> *[85] 'Loomis has not met his burden of showing that the circuit court actually relied on gender as a factor in imposing its sentence. The circuit court explained that it considered multiple factors that supported the sentence it imposed…. In addition to the COMPAS risk assessment, the seriousness of the crime and Loomis's criminal history both bear a nexus to the sentence imposed.'*

Secondly, the Court held that any use of gender in the COMPAS algorithm had a legitimate factual basis:

> *[83] 'there is a factual basis underlying COMPAS's use of gender in calculating risk scores. It appears that any risk assessment tool which fails to differentiate between men and woman will misclassify both genders. As one commenter noted, "the failure to take gender into consideration, at least when predicting recidivism risk, itself is unjust." Melissa Hamilton, Risk-Needs Assessment: Constitutional and Ethical Challenges, 52 Am. Crim. L. Rev. 231, 255 (Spring 2015). Thus, if the inclusion of gender promotes accuracy, it serves the interests of institutions and defendants, rather than a discriminatory purpose.'*

The Court went further and positively approved of COMPAS's use of gender on the basis that it improved the accuracy of criminal sentencing:

> *[86] 'We determine that COMPAS's use of gender promotes accuracy that ultimately inures to the benefit of the justice system including defendants.*

Non-disclosure of COMPAS

At no stage did the Supreme Court, or the sentencing judge, order that the commercial owners of COMPAS disclose the technical specifications of the algorithm to Mr Loomis or to the Court. Indeed, the Supreme Court rejected an offer by COMPAS's commercial owners offered to give evidence about the operation of its algorithm.

The absence of any specific information regarding the operation of COMPAS undermined the Court's decision to dismiss Mr Loomis's appeal. Each appeal ground (accurate information, individualised sentence and sex discrimination) challenged the process of mechanical reasoning employed by COMPAS in assigning recidivism risk scores, and transparent and credible reasoning about each of grounds relied on engagement with the code and datasets which supported COMPAS.

A judge of the Supreme Court acknowledged this significant problem:

> *[132] 'this court's lack of understanding of COMPAS was a significant problem in the instant case. At oral argument, the court repeatedly questioned both the State's and defendant's counsel about how COMPAS works. Few answers were available.'*

Additionally, in providing advice on the future use of COMPAS in sentencing, the Supreme Court recommended that prosecutors must inform sentencing judges of the following "cautions regarding a COMPAS risk assessments accuracy":

> *(1) the proprietary nature of COMPAS has been invoked to prevent disclosure of information relating to how factors are weighed or how risk scores are to be determined;*
>
> *(2) risk assessment compares defendants to a national sample, but no cross validation study for a Wisconsin population has yet been completed;*
>
> *(3) some studies of COMPAS risk assessment scores have raised questions about whether they disproportionately classify minority offenders as having a higher risk of recidivism; and*
>
> *(4) risk assessment tools must be constantly monitored and re-normed for accuracy due to changing populations and subpopulations.*

Importantly, none of those cautions prevent a sentencing judge from giving heavy weight to COMPAS in sentencing a person to imprisonment, despite knowing nothing about the technical specifications of COMPAS.

Audit standards

Knowledge: no compliance (0/4)

There was no compliance with the Knoweldge Criterion.

The algorithm that powered COMPAS was (and remains) a trade secret. People whose data may be fed into COMPAS for the purpose of a risk assessment are completely unable to determine how the machine learning technology which underlies COMPAS operates. In those circumstances, it is impossible for ordinary people to know how COMPAS assigns them recidivism risk scores.

Assent: no compliance (0/4)

There was no compliance with the Assent Criterion.

No legislative regime authorised the use of COMPAS by judges in the Wisconsin criminal justice system. No democratic consent to the use of COMPAS was expressed through legislation. Thus, ordinary people had not assented to the specific use of machine learning technology before it was used by a government official.

Personhood: weak compliance (1/4)

There was only weak compliance with the Personhood Criterion.

COMPAS was acknowledged by its developers to be informed by group information – i.e. information that was about people like Mr. Loomis, rather than Mr. Loomis himself. Though judicial discretion remained at the level of the final decision-making stage, this is an extraordinary shift away from a justice standard oriented at a person's own behaviour in the world and the consequences that should follow for them.

Basic Protections: weak compliance (1/4)

There was only weak compliance with the Basic Protections Criterion in the *Loomis* case.

Although US (federal and State) legislation provided anti-discrimination and due process laws they were ineffective.

As subsequent studies have established, COMPAS produced materially different outcomes depending on the race and gender of the person it was assessing. The impact of those studies was known to the Wisconsin and Federal judiciary. Those data indicated a clear violation of US anti-discrimination laws, but the judiciary did not tailor those laws to the particular challenges of data-drive machine learning technologies, like COMPAS.

Contestability: weak compliance (1/4)

There was only weak compliance with the Contestability Criterion in the *Loomis*.

While *Loomis* was entitled to contest the legality of the use of COMPAS in his sentencing, the judges who ruled on the legality of the use of COMPAS did not appear to possess a high level of interest in the technical of the COMPAS algorithm. That lack of interest could stem from two factor: a lack of education in the technical details of COMPA

SS; and the failure to order that those technical details be disclosed to the court and Mr Loomis's legal team.On either basis, the absence of any meaningful judicial analysis of the technical details of COMPAS hampered Mr Loomis's capacity to contest the legal of AI in his sentencing.

Remedial Action: weak compliance (1/4)

Finally, there was only weak compliance with the Remedial Action Criterion.

In one sense, the adequacy of any remedies available to Mr Loomis is irrelevant given the court's decision that the use of COMPAS did not breach his legal rights. That understanding of the role of remedies in *Loomis* is unduly narrow because it overlooks the failure of the court to order the disclosure and publication of the technical details of COMPAS.

The absence of any remedial avenue for Mr Loomis to obtain information about the technical foundation of COMPAS was a significant failure of the remedial framework in which the legality of AI in criminal sentencing was administered.

<u>Total Score: 4/24</u>

Audit Criterion	**(0) No compliance**	**(1) Weak compliance**	**(2) Moderate compliance**	**(3) Strong compliance**	**(4) Excellent compliance**
Citizen Knowledge	X				
Assent	X				
Personhood		X			
Basic Protections		X			
Contestabilit y		X			
Remedial Action		X			

Comment on comparable systems

The low score of US law regarding the use of COMPAS would likely be replicated in other jurisdictions.

Importantly, the use of machine learning software (like COMPAS) in sentencing is not prohibited under the GDPR. Such use would not fall within the discrete rules concerning 'Automated decision-making, including profiling' in Art 22 of the GDPR: 'The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her,' (See the developing literature on Art 21/22: Kaminski 2019; Dreyer and Schulz 2019). That legal rule would not prohibit sentencing judges from using COMPAS-like technologies in *Loomis*-like situations because the final decision to sentencing would remain with the judge, circumventing the 'decision based solely on automated processing' norm from Art 22 of the GDPR.

Other jurisdictions may impose more robust requirements to disclose/explain the technical specifications of machine learning systems similar to COMPAS, potentially including the code and data sets which underpin the creation of a recidivism risk assessment. A prominent example can be found in Australia, where a judge refused to rely on an algorithmic risk assessment in refusing to order the continued detention of an indigenous Australian person serving a sentence for serious sexual offences (See: Director of Public Prosecutions for Western Australia v. Mangolamara 2007; Stobbs et al. 2017). The judge's critical attitude towards the use of algorithmic risk-assessment tools contrasts sharply with the light-touch adopted by the US judiciary in *Loomis*:

> *165 In the end, bearing in mind that the rules of evidence reflect a form of wisdom based on logic and experience, I am of the view, for the reasons I have referred to, that little weight should be given to those parts of the reports concerning the assessment tools. In my view, the evidence in question does not conform to long-established rules concerning expert evidence. The research data and methods underlying the assessment tools are assumed to be correct but this has not been established by the evidence. It has not been made clear to me whether the context for which the categories of assessment reflected in the relevant texts or manuals were devised is that of treatment and intervention or that of sentencing. Dr Pascu acknowledged under cross-examination that the assessment tools are directed not to the commission of serious sexual offences but to sexual re-offending of any kind (t/s 60). She acknowledged also that the database used for the mathematical model upon which Static-99 was based related to untreated English and Canadian sex offenders released back into the community on an unsupervised basis (t/s 68).*
>
> *166 Moreover, having regard to the admissions made under cross-examination that the tools were not devised for and do not necessarily take account of the*

social circumstances of indigenous Australians in remote communities, I harbour grave reservations as to whether a person of the respondent's background can be easily fitted within the categories of appraisal presently allowed for by the assessment tools.

Data Archiving/networking

The next type of law to be audited applies to the use of data archiving/networking technologies by public sector agencies.

As Part I explained, data archiving/networking technology are digital systems capable of archiving very large amounts of information and transmitting that information through digital networks. While those systems can integrate automation and machine learning technologies, they ultimately rely on a discrete set of technologies:

- *Digitisation*: coding of information in machine-readable format;
- *Data archiving*: storage and ordering of information in large data sets; and
- *Digital networking*: connecting many digital computers to those data sets.

The case study selected to audit the compliance of legal regimes governing data archiving/networking is drawn from the transfer of medical data from the UK National Health Service (NHS) to a Google subsidiary.

Google DeepMind

In 2015, a major UK public health authority transferred around 1.6 million people's patient records to a subsidiary of Google/Alphabet: DeepMind Technologies Ltd (**DeepMind**). Although the maintenance of those patient records by the NHS was perfectly lawful, the subsequent transfer of whole-of-hospital data troves to DeepMind (via Google) was without lawful basis and, in 2017, the UK Information Commissioner's Office ultimately found that it involved five breaches of the Data Protection Act (UK) (See, generally: "Google DeepMind and healthcare in an age of algorithms").

That vast transferral of confidential health records occurred as a result of a data sharing agreement between the NHS and Google Ltd (Google's UK division). In 2016, the scale of the data transfer was disclosed: fully-identifiable patient data, which was not limited to either the data set or the patient group which had been claimed for the transfer, namely, those suffering from acute kidney injury (AKI), which DeepMind proposed to assist by developing a clinical alert app (based exclusively on automation technology, not machine learning). It was later revealed that the NHS and DeepMind did not consult any relevant public bodies (Information Commissioner, Health Research Authority, or Medicines and Healthcare products Regulatory Agency) when they entered into the data sharing agreement. In 2017, the UK's Information Commission ruled that the NHS failed to comply with the relevant UK data protection legislation.

Technology

The transfer of patient data from the UK's public health service to Google was facilitated by data archiving/networking technologies.

In many respects, the technology underlying the storage and transmission of NHS patient data in the DeepMind case would be intuitively familiar to many people. Patient records were digitised from the point of entry into the NHS records via computer programs for the storage of plain text and digital image files. Those digitised records were stored in (on or off-site) NHS data archives. The data stored in those achieves could be transmitted from the NHS to DeepMind via a Wide-Area-Network (i.e. the Internet) using various types of file transfer protocols which provided for the encryption of patient data between NHS archives and Google's servers.[30]

Impact on ordinary people

The use of data archiving/networking to transfer patient details from the NHS to DeepMind benefited patients with AKI by providing effective clinical alerts. For other patients, however, which comprised some 5/6 of the data set, there was no clinical purpose for the transfer.

The transmission of vast quantities of identifiable highly-sensitive personal health data without consent or express purpose has a number of concrete negative impacts, including psychological harms stemming from violations of trust and privacy, as well as potential reputational and financial impacts. Additionally, the transmission of confidential health data to a for-profit company constituted the transfer of valuable economic resources from individuals (via the NHS) without the provision of any compensation or protection.

Legal rules

Two overlapping legal regimes governed the NHS-DeepMind data transfers: data protection law, health care records law and human rights law.

The first was the *Data Protection Act 1998* (UK) which was built on EU law and imposed several basic obligations regarding NHS patient data. First, the NHS was generally required to seek consent of each patient before transferring their health data to another entity. Secondly, the NHS was permitted to transfer data without consent if the transferral was for "medical purposes" which included "preventative medicine" undertaken by a "medical professional" or a person owing duties of confidentiality "equivalent…to a health professional".[31] Thirdly, DeepMind could only use the patient data for its business purposes with the explicit patient consent or to undertake preventative medicine. DeepMind maintained that it was in a direct care relationship, carrying with it implied consent, with every single patient in the hospital as a justification for the transfer. Various enforcement mechanisms are provided by the Data

[30] Streaming via TCP/IP encrypted channel, SSH File Transfer Protocol, N3 and/or AES 256bit encryption.

[31] *DPA 1998, ss 1, 2, 4 and Sch 8*.

Protection Act, though the most prominent of those, requiring a person to cease processing data or an offence (punishable by a fine) of unlawfully processing data, were not applied to the case, and the ICO only considered the behaviour of the NHS trust, rather than DeepMind.

The second legal regime governing the NHS-DeepMind transfers was contained in legislation specifically relating to the use of patient records by the NHS: the *National Health Service Act 2006* (UK) and *The Health Service (Control of Patient Information) Regulations 2002* (UK). Together those legislative instruments permit a person to apply to the UK Secretary of State for Health to obtain and use patient data without consent in a very limited set of circumstances, relevantly including 'diagnosing' ,'recognising trends in' ,'controlling and prevent the spread of' 'communicable diseases'.[32] If permission is granted to use patient data in that way, the entity using the data must actively assist the Health Secretary in investigating and auditing the use of patient data. No such application was made in the DeepMind case.

Law in operation

The NHS-DeepMind data transfers occurred without any notice or consent from any patients. Nor were any regulators consulted regarding the transfers. All of these omissions were in breach of the *Data Protection Act 1998* (UK).

The NHS-DeepMind data transfers were never subject to formal legal challenge in the courts. Instead, based on prominent efforts of journalistic and academic investigation, various regulatory bodies were spurred to action, some of which commenced formal investigations into the legality of the data transfers. The most notable investigation was that conducted by the ICO, which concluded that the Data Protection Act had been breached. Despite that finding, no mandatory enforcement action was taken against the NHS or DeepMind: instead, the ICO requested that the NHS voluntarily agree not to breach UK laws in the same way again in the future. No action was taken to remove the unlawfully transferred data from DeepMind's custody: instead, the ICO expressly permitted DeepMind to continue using the patient data (See: "Letter from Information Commissioner to NHS" (3 July 2017); Powles 2017).

Critically, despite the obvious lack of a legal basis for the vast transfer of confidential patient data, no judicial proceedings were invoked, no enforcement action was taken against NHS or DeepMind, leaving the legal pre-conditions for legally using of health data (consent, legitimate purpose, and other data protection principles) entirely unenforced.

[32] Reg 3.

Audit Criteria

Knowledge: weak compliance (1/4)

Unlike automation and ML, a high level of public knowledge about archiving and networking can be assumed by ordinary members of society given the ubiquity of personal use of digital computing and the internet. In that sense, there is a much higher-degree of knowledge in the general public concerning the technical means through which the NHS-DeepMind transfers occurred.

However, it is far less clear that the use of data archiving and networking would be used to connect public health record systems with the servers of for-profit companies which have no health-care expertise or track-record. It is highly-contestable whether ordinary members of the public would have knowledge that health data typed into their doctor's desktop computer could be transmitted to a Google subsidiary company without their knowledge or prior consent.

For that reason, there is only weak compliance with the Knoweldge Criterion.

Assent: no compliance (0/4)

There was no compliance with the Assent Criterion.

The use of data archiving and transmission systems to bulk transfer patient records (without patient knowledge or consent) was never expressly authorised by legislation.

Nothing about that position is altered by the existence of various "consent" requirements in the UK data protection and health records legislation. Those requirements were premised on direct care, which did not authorise the use of the data archiving/networking technologies which facilitated the bulk transfer of NHS patient records to DeepMind.

Personhood: weak compliance (1/4)

There was weak compliance with the Personhood Criterion.

Data was transferred on every patient, rather than for particular patients with medical care requirements relating to acute kidney injury. In fact, as experts commented at the time, there were many more patients in the data set who were *not* suffering from AKI compared to those who were. The transfer also concerned people who were no longer patients at the hospital, and even patients who were no longer alive.

The systems ultimately developed as a result of the transfer also posed significant risks to the personhood standard. Again, as commentators at the time observed, the need for an AKI detection algorithm was in many ways a replacement for nurses ensuring that patients on wards were well hydrated.

Basic protections: weak compliance (1/4)

There was weak compliance with the Basic Protections Criterion.

The critical right in need of protection was privacy of confidential health records which could only be qualified by explicit consent and legitimate medical care. Although UK law provided formal legal protection of that right in the *Data Protection Act* and *National Health Service Act*, there was no satisfactory institutional mechanism backing up the letter of the law.

The contractual agreement to transfer the patient data from the NHS to Google/ DeepMind was executed, and data transfer commenced, without any independent legal oversight. While that was certainly a governance failure, it also indicated a failure of legal design: the fact that vast troves of highly sensitive personal data were transferred without any meaningful legal justification carried no obvious legal consequences for the public health authority, the public health officials and the private company which facilitated the transferral.

Contestability: weak compliance (1/4)

The legal framework governing the NHS-DeepMind transfers only provided weak compliance with the Contestability Criterion.

Institutions for the enforcement of consent and medical treatment principles were weak.

The Information Commissioner's powers were strong on paper, but weak in practice. The Commissioner had strong legal powers to investigate unlawful data transfers, to order that unlawfully transferred data be deleted or returned and to assist and notify patients affected by unlawful data transfers. None of those powers were exercised, despite the patent unlawfulness of the NHS-DeepMind data transfers.

There was also a theoretical possibility that the NHS or DeepMind could be sued by individual patients for harmful or unlawful use of their confidential health data. Practical obstacles to that type of contest were virtually insurmountable. Individual patients had no way of knowing that their health records were transferred to a Google subsidiary until after transfers had already occurred. Even if that knowledge were obtained, vast financial and social obstacles stood in the way of a single patient commencing proceedings against the UK's public health authority and a subsidiary of a +3 global tech company.

Remedial Action: weak compliance (1/4)

There was weak compliance with the Remedial Action standard.

Legal rules existed for the deletion of unlawfully obtained data, the prosecution of people who obtained that data and compensation of people harmed by unlawful data transfers existed. However, each of those rules relied on exceptionally weak institutional mechanisms for enforcement.

Total Score: 5/24

Audit Criterion	(0) No compliance	(1) Weak compliance	(2) Moderate compliance	(3) Strong compliance	(4) Excellent compliance
Citizen Knowledge		X			
Assent	X				
Personhood		X			
Basic Protections		X			
Contestability		X			
Remedial Action		X			

Comment on comparable systems

The low score of UK legislation in relation to health data and its unlawful transfer to third parties would likely be replicated in other jurisdictions. Many major privacy and health data legislative regimes provide the same two basic principles for the transferral and use of health data: patient consent or the provision of health care services.

Most major systems rely

on the same institutional structure to enforce those standards: independent statutory offices (data protection authorities), informed by medical care guidelines, charged with providing both oversight and guidance to institutions processing data, and with little institutional heft to provide effective enforcement.The GDPR provides the same basic principles for the lawful use of health data as appeared in the UK *Data Protection Act 1998*: consent and the purpose of “ensuring high standards of quality and safety of health care.” The precise application and meaning of those standards will vary jurisdictionally, but a common factor is the relatively weak position of data protection authorities compared to the enormous institutional and economic authority of core governmental departments and tech majors.

Mass Surveillance

The final type of law to be audited applies to the use of mass surveillance technologies by governments.

As Part I explained, mass surveillance technologies permit the instantaneous observation, recording and storage of information concerning individual human behaviour and human social interactions, including:

- Voice;
- Text (whether hardcopy or softcopy);
- Images (including facial recognition);
- Biometrics (biological information unique to a single human being, as well as inferred from populations); and
- Geolocative data.

While mass surveillance can integrate automation, machine learning and data archiving/ networking, it essentially relies on hardware technology of the following kinds:

- Cameras;
- Biometric scanners: fingerprint, voice, retinal, facial, gait, body;
- Mobile computing devices and applications: laptop/desktop computers, tablets, smartphones;
- Physical access points for those cameras, scanners and devices; and
- Networks which permit transmission of information collected from those cameras, scanners and computing devices to storage facilities.

The case-study selected to audit the compliance of legal regimes governing mass surveillance is drawn from the use of live facial recognition technology in the UK in 2017.

Facial recognition in the UK

In 2017, a UK police force began using facial recognition technology to identify and locate suspected and convicted criminals in public areas (See: South Wales Police, "Facial Recognition Technology"; "NeoFace Watch"). The specific technology was provided by NeoFace Watch Inc, a US company which sold facial recognition technology to public and private sector entities.
That technology was deployed as form of 'live facial recognition': using CCTV camera images to match the faces of people in public spaces to databases of suspected/ convicted criminals archived in the system. The technical details of that form of facial recognition are explained below.

The police used live facial recognition at a busy shopping centre in 2017 and at a military equipment expo in 2018. A privacy advocate, Mr Edward Bridges, attended both events and claimed that the police used that technology to capture his facial image unlawfully.

Mr Bridges sued the police force under the UK's human rights, anti-discrimination and data protection legislation.

Technology

The facial recognition system developed by NeoFace Watch had five key components.

1. **Data collection**: first, a large data set of facial images of suspected or convicted criminals was complied: the "Watchlist". Once compiled, facial images were processed by optical recognition software so their features could be quantified and used to provide high-speed matching with new facial images.

2. **Face-capture cameras**: secondly, CCTV cameras take high-resolution digital photographs of people in public (or private) spaces. The facial features of the photographed people were then extracted and quantified for matching with facial-feature data contained in the Watchlist. The resulting data is "biometric data": ie, biological data about a person which can be analysed by computative systems, including by identifying the person.

3. **Facial-feature matching**: the biometric data captured through CCTV cameras was then compared to the facial-feature data in the Watchlist using quantitative methods which produce results indicating whether there is a "match" of the biometric data captured through CCTV and the Watchlist. The resulting output is probabilistic, because NeoFace Watch outputs a "similarity score" which indicates the likelihood of a positive match, rather than a certainty of a positive match.

4. **Flagging or deletion**: once NeoFace Watch has determined the degree of match, it provides its human operators with several options, including: (i) to identify the person captured on CCTV as a person on the Watchlist due to a high similarity score; (ii) to retain the captured biometric facial data of that person; (iii) delete the captured biometric facial data because of a low similarity score.

5. **Human review**: a human operator is presented with the two faces (the face captured via CCTV and its match in the Watchlist) and the decision is made whether to exercise police powers to stop, interrogate and detain the identified person.

Impact on ordinary people

The facial recognition system provided by NeoFace Watch has obvious advantages for law enforcement bodies and the general public. It provides enormous time and resource efficiencies in surveillance, facilitating the rapid apprehension of people wanted for actual or suspected crimes. In that sense, NeoFace Watch has a beneficial impact on ordinary citizens: protecting them from physical, emotion and economic harm, and reducing fear in communities affected by crime.

NeoFace Watch does, however, have negative impacts. Through bulk collection and quantification of people's biometric facial data, NeoFace Watch impinges upon people's privacy and their right to a private life. When deployed by government (particularly law enforcement) agencies, that impingement on people's private lives leads to a chilling of legitimate activities of intellectual, cultural or emotional disagreement, debate, protest and dissent. Additionally, potential biases built into the algorithmic design of NeoFace Watch expose people in certain cultural and biological groupings to discrimination, via erroneously high match-rates of captured biometric data to facial-feature data in the Watchlist.

Legal rules

The legality of NeoFace Watch was governed by 4 distinct legal regimes in the UK.

First, and most basically, NeoFace Watch was governed by legislative and judge-made law which empowered police officers to undertake inquiries and surveillance to prevent crime and keep the peace. UK police are under a legal duty to prevent and detect crime, and have legal powers to use, retain and disclose images of people, including by compiling watchlists, necessary to discharge that duty.[33]

Those legal norms provided blank-cheque authorisation to use facial recognition systems. Unless police are required to enter private property, no warrants or notifications were required to capture facial images

Secondly, NeoFace Watch was governed by the human rights law contained in the UK *Human Rights Act 1998*. The most immediately relevant human right challenged by the use of NeoFace Watch was the right to a private life, free from disproportionate government interference:

[33] *Rice v Connolly* [1966] 2 QB 414 at 419B – C; *R (Catt) v Association of Chief Police Officers* [2015] AC 1065 at [7]; *Police and Criminal Evidence Act 1984* (UK), s 64A.

> 1. Everyone has the right to respect for his private and family life, his home and his correspondence.
> 2. There shall be no interference by a public authority with the exercise of this right except such as is in accordance with the law and is necessary in a democratic society in the interests of national security, public safety or the economic wellbeing of the country, for the prevention of disorder or crime, for the protection of health or morals, or for the protection of the rights and freedoms of others.

Thirdly, the UK's data-protection legislation (modelled on the EU's GDPR) governed the collection and use of data about people's faces through NeoFace Watch: the *Data Protection Act 2018* (UK) (***DPA 2018***).[34] Relevantly, the *DPA 2018* imposed two key obligations on the use of NeoFace Watch by UK Police.

The first obligation required compliance with core "data protection principles" (s 34 and 35).

> **35 The first data protection principle**
> (1) The first data protection principle is that the processing of personal data for any of the law enforcement purposes must be lawful and fair.
> (2) The processing of personal data for any of the law enforcement purposes is lawful only if and to the extent that it is based on law and either—
> (a) the data subject has given consent to the processing for that purpose, or
> (b) the processing is necessary for the performance of a task carried out for that purpose by a competent authority.
> (3) In addition, where the processing for any of the law enforcement purposes is sensitive processing, the processing is permitted only in the two cases set out in subsections (4) and (5).
> (4) The first case is where—
> (a) the data subject has given consent to the processing for the law enforcement purpose as mentioned in subsection (2)(a), and
> (b) at the time when the processing is carried out, the controller has an appropriate policy document in place.
> (5) The second case is where—
> (a) the processing is strictly necessary for the law enforcement purpose,
> (b) the processing meets at least one of the conditions in Schedule 8, and
> (c) at the time when the processing is carried out, the controller has an appropriate policy document in place.

> **Schedule 8 Conditions for sensitive processing under Part 3**
> ***Statutory etc purposes***
>
> 1 This condition is met if the processing—
> (a) is necessary for the exercise of a function conferred on a person by an enactment or rule of law, and
> (b) is necessary for reasons of substantial public interest.
> ***Administration of justice***
>
> 2 This condition is met if the processing is necessary for the administration of justice.

[34] Until 2018, an earlier (pre-GDPR) data protection statute governed the use of NeoFace Watch: the *Data Protection Act 1998* (UK).

The second obligation imposed by the *DPA 2018* on UK police's use of NeoFace Watch concerned the carrying out of a "data protection impact assessment" where data processing carries a "high risk to the rights and freedoms of individuals" (s 64).

> **64 Data protection impact assessment**
>
> (1) Where a type of processing is likely to result in a high risk to the rights and freedoms of individuals, the controller must, prior to the processing, carry out a data protection impact assessment.
>
> (2) A data protection impact assessment is an assessment of the impact of the envisaged processing operations on the protection of personal data.
>
> (3) A data protection impact assessment must include the following—
>
> (a) a general description of the envisaged processing operations;
>
> (b) an assessment of the risks to the rights and freedoms of data subjects;
>
> (c) the measures envisaged to address those risks;
>
> (d) safeguards, security measures and mechanisms to ensure the protection of personal data and to demonstrate compliance with this Part, taking into account the rights and legitimate interests of the data subjects and other persons concerned.
>
> (4) In deciding whether a type of processing is likely to result in a high risk to the rights and freedoms of individuals, the controller must take into account the nature, scope, context and purposes of the processing.

Fourthly, Neoface Watch was governed by the *Equality Act 2010* (UK), particularly the "Public Sector Equality Duty" (**PSED**) contained in s 149.

> ***149 Public sector equality duty***
>
> (1) A public authority must, in the exercise of its functions, have due regard to the need to—
>
> (a) eliminate discrimination, harassment, victimisation and any other conduct that is prohibited by or under this Act;
>
> (b) advance equality of opportunity between persons who share a relevant protected characteristic and persons who do not share it;
>
> (c) foster good relations between persons who share a relevant protected characteristic and persons who do not share it.

Law in operation

The legality of NeoFace Watch's automated facial recognition (**AFR**) system was challenged in litigation against an arm of the UK's police force. Three distinct legal arguments were made.

First, NeoFace Watch violated people's human rights to a private life because AFR was not regulated by legal frameworks which provided clear and specific rules governing its use by the UK police agency. Secondly, using NeoFace Watch breached UK data protection legislation because the UK police agency did not have an appropriate policy document in place regulating the deployment of AFR and had failed to issue a Data Protection Impact Assessment which flagged the human rights threats of using AFR. Thirdly, the UK police agency breached their statutory duty to ensure equality and protect against discrimination because the agency failed to undertake continuing investigations into whether NeoFace Watch's AFR system discriminated against people on the basis of sex, gender, race or other legally protected attributes.

Those challenge were eventually upheld by the Court of Appeal of England and Wales (See: R (Bridges) v Chief Constable of South Wales Police [2020]).

Right to a private life

The Court of Appeal held that the use of NeoFace Watch breached people's right to a private life enshrined in Art 8 of the European Convention on Human Rights and the UK's *Human Rights Act*. That breach arose from the absence of a legal framework surrounding the deployment of AFR that was "compatible with the rule of law", being "accessible" and "foreseeable". Under UK (and European) human rights law, the right to a private life can only be qualified by legislative rules which comply with those dual requirements. The Court recounted that the accessible and foreseeable standards required that legislation governing AFR must meet the following requirements [55]:

> *"The legal basis [of AFR] must be 'accessible' to the person concerned, meaning that it must be published and comprehensible, and it must be possible to discover what its provisions are. The measure must also be 'foreseeable' meaning that it must be possible for a person to foresee its consequences for them and it should not 'confer a discretion so broad that its scope is in practice dependent on the will of those who apply it, rather than on the law itself."*
>
> *the law must 'afford adequate legal protection against arbitrariness and accordingly indicate with sufficient clarity the scope of discretion conferred on the competent authorities and the manner of its exercise"*

It was common ground that no legislation specifically authorised the use of NeoFace Watch, or AFR generally, by the UK police. In the absence of such legislation, the counter-argument was raised that policy documents issued by the Surveillance Camera Commissioner and the UK police were "laws" which met the accessibility and foreseeability standards. The Court rejected those arguments holding that

> *[91] "too much discretion is currently left to individual police officers. It is not clear who can be placed on the watchlist nor is it clear that there are any criteria for determining where AFR can be deployed.*
>
> *[94] We are satisfied…that the current policies do not sufficiently set out the terms on which discretionary powers can be exercised by the police and for that reason do not have the necessary quality of law."*

Data protection claims

The Court held that the UK police agency breached the *Data Protection Act 2018* because it failed to issue a Data Protection Impact Assessment (**DIPA**) that assessed the risk to rights and freedoms, specifically the human right to private life. The UK Police Agency had issued a DIPA, but the DIPA had failed to identify the violation of the right to a private life which flowed from the lack of an accessible and foreseeable legislative framework for the use of NeoFace Watch.

Interestingly, the UK judiciary approved the use of AFR under the *Data Protection Act 2018* without any legislative foundation on the ground that the common law powers of constables is a sufficient "basis in law" for the processing of biometric facial data.

Anti-discrimination claims

The Court held that the UK police agency had failed to discharge its Public Sector Equality Duty (**PSED**) because it failed to continuously assess whether NeoFace Watch was designed and operated in such a way that it discriminated against people on the basis of gender, race or other protected attribute.

Both parties led evidence on the question whether NeoFace Watch matched faces at error rates which indicated a discriminatory effect against men/women and people of different races/ethnicities. The UK police agency relied on the evidence of a police constable who reviewed the rate of positive/false matches which were identified during an earlier trial deployment of NeoFace Watch in which 290 alerts were generated:

> *[188] 188 of the alerts were males (65%). Of the 188 male alerts, 64 (34%) were true positives and 124 (66%) were false positives. In relation to females, of 102 alerts, 18 (18%) were true positives and 84 (82%) were false positives. A number of the female false alerts were matched against primarily two individuals who the AFR software provider would refer to as a "lamb". A lamb is a person whose face has such generic features that may match much more frequently.*
>
> *[189.] [the police constable] also reviewed the ethnicity of those who were the subject of an alert. Of the true positives (82) 98% were "white north European". Of the false positives (208) 98.5% were "white north European".*
>
> *[190.] [the police constable] therefore concluded…: "From my experience and the information available to me, I have seen no bias based on either gender or ethnicity.*

The Court emphatically ruled that such evidence did not show compliance with the statutory PSED:

> *[The police constable] did not know, for obvious reasons, the racial or gender profiles of the total number of people who were captured by the AFR technology but whose data was then almost immediately deleted. In order to check the racial or gender bias in the technology, that information would have to be*

> *known. We accept … that it is impossible to have that information, precisely because a safeguard in the present arrangements is that that data is deleted in the vast majority of cases. That does not mean, however, that the software may not have an inbuilt bias, which needs to be tested. In any event [the police constable] is not an expert who can deal with the technical aspects of the software in this context.*

Ultimately, the Court held that the UK police agency never took meaningful steps to check whether NeoFace Watch operated in a discriminatory way: [199] "…[the UK police] have never sought to satisfy themselves, either directly or by way of independent verification, that the software program in this case does not have an unacceptable bias on grounds of race or sex."

That conclusion was arrived at despite the refusal (as in the COMPAS case-study) of NeoFace Watch to disclose the technical specifications of its AFR system:

> *[199] There is evidence…that programs for AFR can sometimes have such a bias. [NeoFace Watch's employees] cannot comment on this particular software but that is because, for reasons of commercial confidentiality, the manufacturer is not prepared to divulge the details so that it could be tested. That may be understandable but, in our view, it does not enable a public authority to discharge its own, non-delegable, [public sector equality] duty…"*

Audit Criteria

Knowledge: weak compliance (1/4)

There was only weak compliance with the Knoweldge Criterion.

By the deployment of NeoFace Watch in 2018, many people can be assumed to have some vague awareness of the existence of facial recognition technology. Additionally, the UK police agency made some efforts to notify people that AFR technology could be used in public spaces. To that extent, there was some (albeit low) level of knowledge about the AFR systems before the deployment of NeoFace Watch.

However, there was no widespread knowledge about the actual technical operation of NeoFace Watch. The technical specifications of the AFR system were not (and still have not been) disclosed for reasons of commercial confidentiality. It was also clear that the public sector employees who used NeoFace Watch misunderstood the precise technical basis of the AFR system, particularly the way that it could impact on people's rights and freedoms.

For that collection of reasons there is only weak compliance with the Knoweldge Criterion.

Assent: no compliance (0/4)

There was no compliance with the Assent Criterion. No legislation specifically authorised the use of AFR by the UK police agency (whether provided by NeoFace Watch or otherwise).

As explained above, that failure had critical ramifications for the legality of NeoFace Watch under human rights legislation: no specific legislation meant that the right to a private life had been breached by the UK police agency without respecting the rule of law.

Personhood: moderate compliance (2/4)

There was only weak compliance with the Personhood Criterion.

The indiscriminate nature of NeoFace Watch, which presumed a state of suspicion for every citizen, was a significant departure from the general premise for law enforcement action, which requires a degree of reasonable cause related to the specific behaviour of individual people.

The law provided some protection against that indiscriminate operation by requiring the existence of accessible and foreseeable legislation in order to comply with the requirements of Art 8 of the ECHR. In that way, there was moderate (albeit indirect) compliance with the Personhood Criterion.

Basic protections: strong compliance (3/4)

The UK law governing the use of NeoFace Watch provided strong protection of basic rights. The central right was the right to private life (or privacy) enshrined in European treaty and UK statute law. The threat to that right presented by AFR was high, but effectively addressed by judicially enforceable human rights law.

Rights embedded in anti-discrimination frameworks were also effectively protected through UK equality legislation and the imposition of the judicially-enforceable Public Sector Equality Duty on the UK police agency.

General privacy law rights were less effectively protected as the UK *Data Protection Act 2018* permitted the collection of biometric facial data through AFR systems without any requirement for consent or a legislative foundation.

Contestability: strong compliance (3/4)

There was also strong compliance with the Contestability Criterion.

Judicial review proceedings in the English and Welsh courts provided an effective institutional mechanism for members of the public to challenge the legality of NeoFace Watch.

The Court of Appeal judges approached the complex technical issues in the case with an awareness of the technical functions of AFR software and its various positive and negative features.

Remedial Action: moderate compliance (2/4)

There was moderate compliance with the Remedial Action Criterion.

The remedy issued by the Court of Appeal in response to the illegal use of NeoFace Watch was a non-coercive declaratory order: announcing that the UK police agency's use of NeoFace Watch violated the human right to a private life, breached the *Data Protection Act 2018* and the Public Sector Equality Duty under the *Equality Act*.

That order represented an important clarification of the legal position regarding AFR, but it did not completely cure the recorded illegality. The Court could have (but did not) order an injunction, which would have forced the UK police agency to cease using NeoFace Watch until it could prove that its use was legal. Nor did the Court make any order regarding deletion or restitution of the biometric data collected through the unlawful use of NeoFace Watch.

Total Score: 11/24

Audit Criterion	**(0) No compliance**	**(1) Weak compliance**	**(2) Moderate compliance**	**(3) Strong compliance**	**(4) Excellent compliance**
Citizen knowledge		X			
Assent	X				
Personhood			X		
Basic protections				X	
Contestability				X	
Remedial Action			X		

Comment on comparable systems

The score given to the law governing facial recognition systems is likely to vary significantly between different legal systems.

A comparably low score for the Knowledge and Assent Criteria can be expected in most jurisdictions. The nature of software and hardware specifications of AFR are almost invariable commercially sen

sitive and therefore kept confidential. There are no examples of national legislation which specifically authorises and regulates the use of AFR. Some sub-national jurisdictions have taken steps to directly regulate the use of facial recognition technology. For example, Washington State in the USA has legislated to regulate the use of facial recognition, although the strength of that legislation has been queried on the basis that the politician with responsibility for drafting and introducing the bill (a senator in the Washington State legislature) was (and is) an employee of Microsoft Corporation (See: "Microsoft Looms Over the Privacy Debate in Its Home State"; "A Microsoft Employee Literally Wrote Washington's Facial Recognition Law").The score for the remaining Criteria (Personhood, Basic Protections, Contestability and Remedial Action) will vary depending on the broader legal and institutional protection afforded to rights to privacy in public places.

In European jurisdictions, a similar outcome to the UK case-study can be expected via the requirements under Art 8 of ECHR requiring accessible and foreseeable legislative rules for the use of AFR. In non-European jurisdictions with established human rights law, the matter is more complicated. For example, neither Canadian nor US human/constitutional rights frameworks contain a explicit right to a private life. Both legal systems provide some privacy based rights protection, but it is radically unclear who those protections would apply to facial recognition.

In jurisdictions without explicit human rights protections there the score is likely to be significantly lower. For example, there would be no obvious legal basis to prevent law enforcement bodies using AFR under Australian law.

Part V: Audit results

	Knowledge	Assent	Personhood	Basic protections	Contestability	Remedial Action	Total
Automation	0/4	0/4	3/4	2/4	2/4	2/4	8/24
Machine learning	0/4	0/4	1/4	1/4	1/4	1/4	4/24
Data archiving/ networking	1/4	0/4	1/4	1/4	1/4	1/4	5/24
Mass surveillance	1/4	0/4	2/4	3/4	3/4	2/4	11/24

Across the case studies, there was no evidence of specific knowledge before deployment of any of the technologies by government. That absence of meaningful knowledge about the technical specifications of the various automation, machine learning, data archiving/networking and mass surveillance technologies is a striking common feature.

Another striking common feature of all the case studies was an absence of compliance with the Assent Criterion. There was no specific legislative authorisation of automation, machine learning, data archiving/networking or mass surveillance technologies. An Australian statute did refer to the authorisation of the automated OCI debt recovery system, however that authorisation was neither released to the public, nor specific in its terms. In the NeoFace Watch case study, the absence of specific legislative authorisation (and regulation) was fatal to the legality of the AFR system under European human rights law.

There was generally weak compliance with the Personhood Criterion across the case studies. A unique feature of each of the legal regimes governing AI was that they failed to explicitly require that governments treat people

as unique individuals, rather than generalising government action towards individuals within cohorts. The strongest compliance with the Personhood Criterion appeared in the Australian OCI case study, where the law directly penalised a failure to tailor government action towards individual circumstances. The weakest appeared in the COMPAS and NHS-DeepMind case studies, where no adverse legal consequences attached to the harmful uses of AI that wholly failed to respect individual autonomy. There was generally a higher level of compliance with the Basic Protections, Contestability and

Remedial Action Criteria across the case studies, indicating that existing legal frameworks governing human rights, privacy and their judicial enforcement are (or, at least, can be) tailored to the unique challenges presented by the use of AI by governments. The outlier is the low scores recorded in the NHS->DeepMind case study, where institutional weakness entirely prevented the enforcement of privacy law.

An interesting overall discovery is that highest total score appeared in the mass surveillance case study, despite this being one of the areas of greatest imbalance in power between state and citizens.