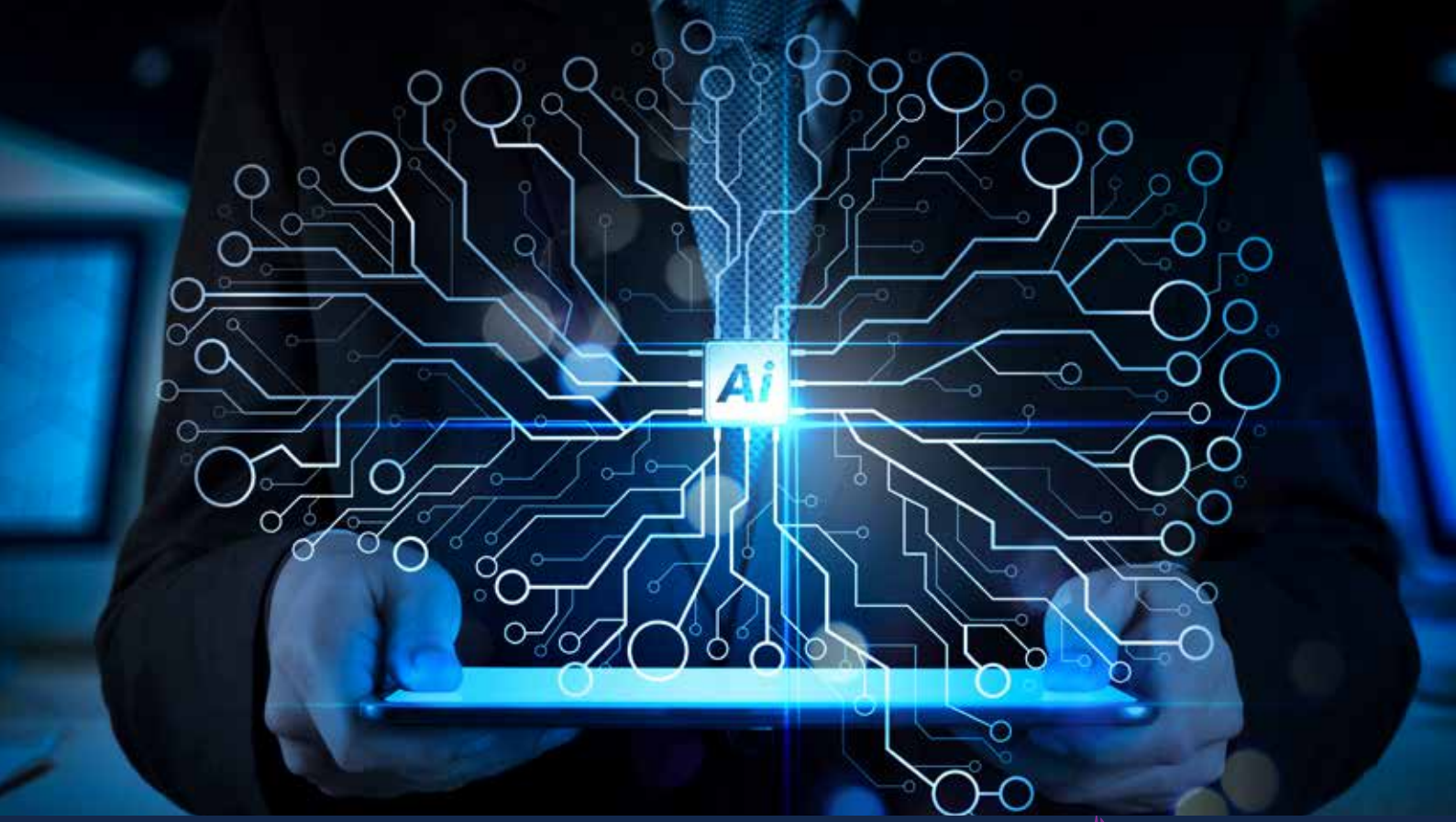
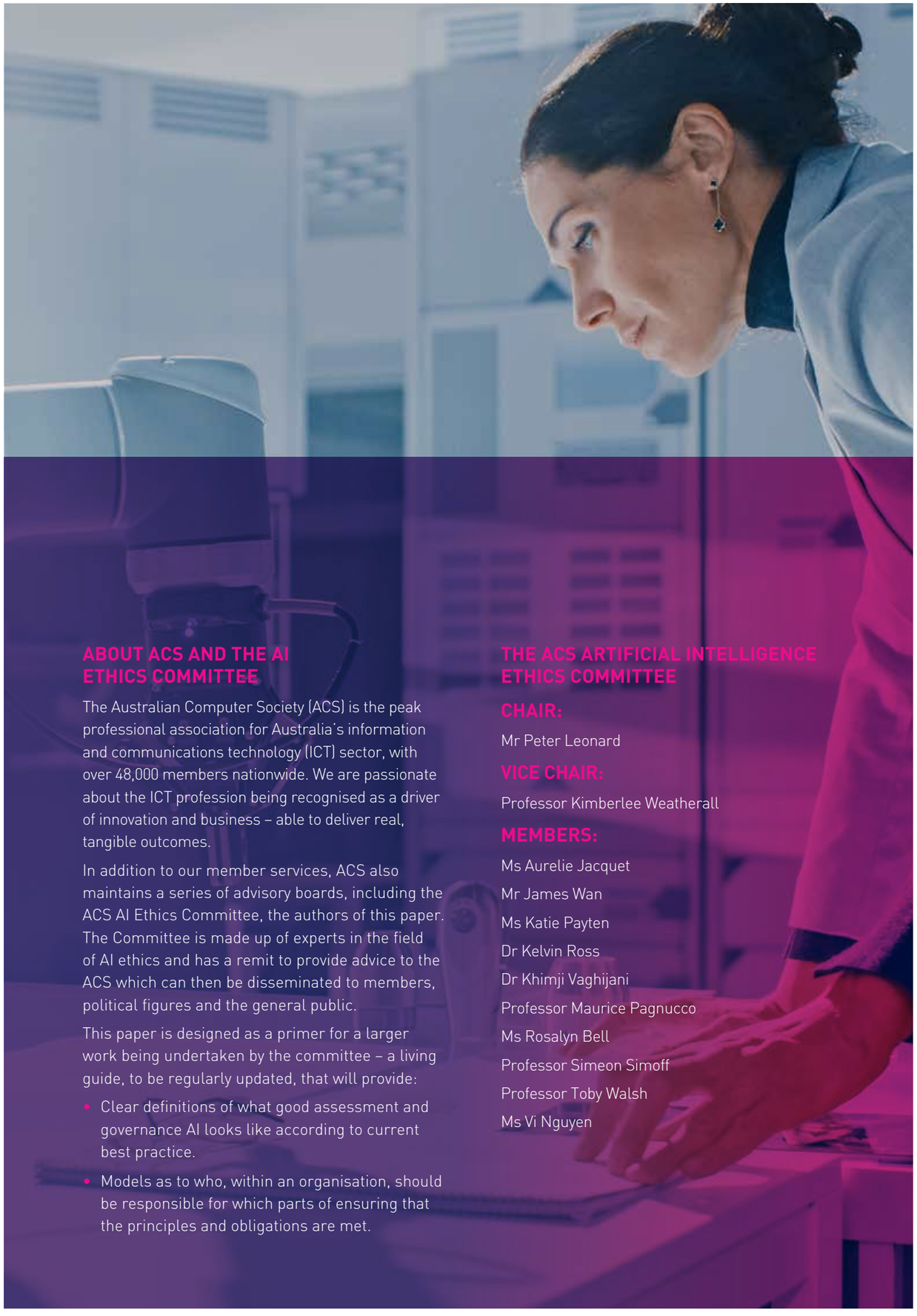


June 2021

The Ethics and Risks of AI Decision-Making

An ACS Primer



ABOUT ACS AND THE AI ETHICS COMMITTEE

The Australian Computer Society (ACS) is the peak professional association for Australia's information and communications technology (ICT) sector, with over 48,000 members nationwide. We are passionate about the ICT profession being recognised as a driver of innovation and business – able to deliver real, tangible outcomes.

In addition to our member services, ACS also maintains a series of advisory boards, including the ACS AI Ethics Committee, the authors of this paper. The Committee is made up of experts in the field of AI ethics and has a remit to provide advice to the ACS which can then be disseminated to members, political figures and the general public.

This paper is designed as a primer for a larger work being undertaken by the committee – a living guide, to be regularly updated, that will provide:

- Clear definitions of what good assessment and governance AI looks like according to current best practice.
- Models as to who, within an organisation, should be responsible for which parts of ensuring that the principles and obligations are met.

THE ACS ARTIFICIAL INTELLIGENCE ETHICS COMMITTEE

CHAIR:

Mr Peter Leonard

VICE CHAIR:

Professor Kimberlee Weatherall

MEMBERS:

Ms Aurelie Jacquet

Mr James Wan

Ms Katie Payten

Dr Kelvin Ross

Dr Khimji Vaghijani

Professor Maurice Pagnucco

Ms Rosalyn Bell

Professor Simeon Simoff

Professor Toby Walsh

Ms Vi Nguyen

Foreword

Peter Leonard
Chair of the ACS Artificial Intelligence Ethics Committee

Artificial intelligence (AI), machine learning (ML) and other automation-assisted decision-making may reduce risks of human error and bias, but may also introduce new and unanticipated risks of errors, unfairness or illegality. Consequences may be rapid and substantial.

Take the Australian Government's Online Compliance Intervention system now better known as Robodebt. Robodebt's automated decision-making generated notices of demand for repayment of welfare payments where the recipient was identified as having received excess benefits. It did that by averaging income over a period, and the averaging period yielded a significant number of false positives, leading to many welfare recipients receiving demands when they had in fact been entitled to benefits for a part year. After years of litigation, the Australian Government in late 2020 agreed to settle for $1.2 billion, made up of $721 million already refunded to more than 370,000 people who had been wrongly pursued; another $112 million in compensation, and a further $398 million in demands being processed that they agreed to drop.

Robodebt illustrates how a calculation that is algorithmically correct when properly applied can be in error, unfair and illegal when applied more broadly, if it's applied without due consideration of errors that can arise and without appropriate human intervention and consideration.

Robodebt is a very public example, but it's just one of many automation-assisted decision-making processes that have blown up as illegal, unfair or biased, causing reputational damage to the organisations that deployed them.

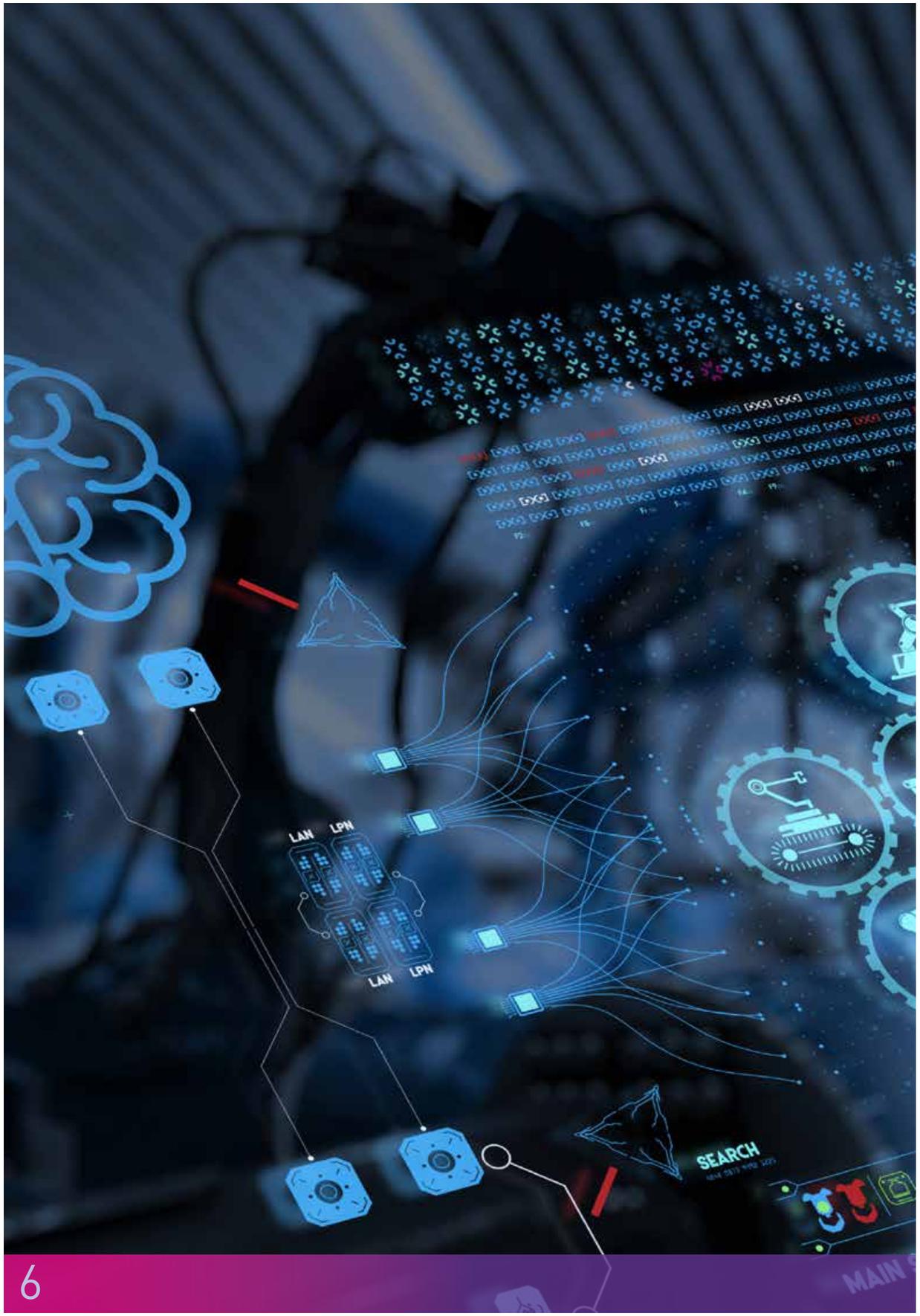
AI, ML and other automation components within decision-making chains require careful consideration in terms of their fairness, accountability and transparency. But many organisations fail to ensure that the automation component is appropriately deployed within a decision-making chain involving people, so that there is not excessive reliance on it or use of it in inappropriate contexts.

At ACS, we've been exploring this issue for a number of years now, looking at the issues of technology and data analytics governance, risk assurance and management, ethics, responsibility to Australian society and the environment, and management frameworks for AI and automation-assisted decision-making. We have more work to present, including work on frameworks and guidance for businesses when it comes to the inclusion of AI agents in decision chains.

To start with we thought we'd provide an overview of this issue – a primer – looking at the issue holistically, as well as some of the efforts, frameworks and legal requirements being developed and considered here in Australia and around the world. The paper you are reading now examines those issues and more.

We'd very much like to thank the members of the AI Ethics Committee for their work on this and the work to come. This is a critical issue as we rely more and more on automation. As a nation and as individuals we need to get this right to ensure that our data privacy, our safety and our economy are not placed in jeopardy by poorly planned deployments and uses of AI systems.

Contents



Introduction

Automated decision-making through artificial intelligence (AI) and machine learning (ML) is being developed and deployed by organisations large and small across all sectors of the Australian economy. This is already delivering many benefits, including improvements in productivity, efficiency, and quality of life, and a capacity to develop new and innovative products and services.

Deployment and use of AI and ML applications also presents new challenges.

Because the risks and the methods to mitigate them are not well understood, new risks could arise that may not be predicted, assessed or managed. This can include people's excessive reliance upon the outputs from the automation applications.

Consequently, assurance of the safety and reliability of decision-making processes in which automation applications are a component requires careful consideration of the interaction between people, technology, data and processes.

This paper looks at some of the challenges and issues created by the inclusion of automated components in the decision-making process. It looks at what makes a 'good' AI, how companies are (and should be) approaching risk management and oversight of decision-making, as well as some of the global regulatory efforts designed to address the issue and ensure the implementation of AI decision-making systems is ethical and in the public interest.

AI, ML AND OTHER AUTOMATION APPLICATIONS

The common feature underlying AI- and ML-enabled decision-making is the use of a combination of data, algorithmic methods and human ingenuity to derive insights and other outputs that are used to influence outcomes.

In this paper, we draw no distinction between advanced data analytics, 'true AI' and ML. Our broader focus is the integration of outputs of data science and automation applications into:

- Human decision-making that affects people or the environment.
- Fully autonomous machine activations; for example, Internet of Things services that cause a machine to act without any direct supervision or control by a human 'in the loop' of the pathway.

01

Criteria for determining whether an automation application is 'good AI'

A great deal of discussion in recent years has focussed upon the need to develop frameworks and methodologies that help data scientists and other information professionals design, build and test automation applications that are reliably fair, ethically accountable and acceptably transparent.

A variety of criteria have been developed for determination of whether an automation application is 'good AI'.[1] Almost all of these elaborate on some version of four basic principles (often reduced to the acronyms FEAT or FATE) that should be applied.

Broadly, the key concepts are:

- *Fairness* to directly impacted individuals as well as the broader societal impact.
- *Ethics* or *equity*.
- *Accountability*.
- *Transparency*, including explainability.

Some of these principles overlap with existing legal obligations under a variety of laws, including in the areas of anti-discrimination, consumer protection, accessibility for persons with disabilities, data privacy, unfair contract terms, misleading and deceptive conduct, product liability and negligence. Accordingly, these principles in part state what is already the law in Australia, and in part state what are evolving societal expectations of good behaviour.

The Australian Government has published AI Ethics Principles,[2] which it summarises as follows:

- Human, social and environmental wellbeing: throughout their lifecycle, AI systems should benefit individuals, society and the environment.
- Human-centred values: throughout their lifecycle, AI systems should respect human rights, diversity and the autonomy of individuals.
- Fairness: throughout their lifecycle, AI systems should be inclusive and accessible, and should not involve or result in unfair discrimination against individuals, communities or groups.
- Privacy protection and security: throughout their lifecycle, AI systems should respect and uphold privacy rights and data protection and ensure the security of data.
- Reliability and safety: throughout their lifecycle, AI systems should reliably operate in accordance with their intended purpose.
- Transparency and explainability: there should be transparency and responsible disclosure to ensure people know when they are being significantly impacted by an AI system, and can find out when an AI system is engaging with them.
- Contestability: when an AI system significantly impacts a person, community, group or environment, there should be a timely process to allow people to challenge the use or output of the AI system.
- Accountability: those responsible for the different phases of the AI system lifecycle should be identifiable and accountable for the outcomes of the AI systems, and human oversight of AI systems should be enabled.

At this time, there is no accompanying regulation in Australia to assure that principles for 'good AI' are given effect in development and use of new technologies. As explored later in this paper:

- Some jurisdictions are considering implementation of legally enforceable obligations for organisations to give effect of some of these principles, at least in the context of development and use of AI systems that are considered higher risk.
- Some jurisdictions are considering changing their legal frameworks to require algorithmic or AI impact assessments in scenarios where the use of automation applications can present significant risks of harms to individuals or the environment.

In essence, proposals for legally mandated impact assessments take existing concepts of privacy impact assessment (in relation to handling of personal information about individuals) and environmental impact assessment, and suggest similar requirements for organisations to evaluate risks and determine appropriate mitigation strategies when commissioning, designing and implementing new automation applications.

There is now broad acceptance of the need to ensure that automation applications are demonstrably and reliably fair, accountable and transparent (sufficiently explainable) in and of themselves; that is, assurance of 'the AI box' itself. We are, however, still early in the process of working out *when* automation impact assessment should be conducted and *how* impact assessments should be conducted. This important work needs to continue.

We need to supplement that work with governance and assurance work to ensure that insertion of automation applications into the decision-making processes within organisations does not lead to other risks. For example, bad outcomes from the use of automation applications may flow from inappropriate or excessive reliance by people on 'the AI box' or the data that feeds it, and not from the inherent design and operation of 'the AI box' itself.

Frameworks and methodologies for governance and assurance must address both the automation application component and the broader decision-making context and environment in which an automation application is used.

Diversity of contexts of use is one challenge that needs to be addressed: it is often difficult for an AI developer to know when and how an application will be used.

Another challenge is the range of capabilities of organisations that are deploying and relying upon automation applications. Australia and other societies must ensure that frameworks for governance and assurance of automation applications are well articulated, well understood and consistently implemented across a diverse range of organisations with differing capabilities and experience in management of business risk. This includes startups, large corporations, charities, government agencies and other organisations.

[1] See *AlgorithmWatch, AI Ethics Guidelines Global Inventory* as updated at https://algorithmwatch.org/en/project/ai-ethics-guidelines-global-inventory/; also *AlgorithmWatch, "Ethical AI guidelines": Binding commitment or simply window dressing?* at https://algorithmwatch.org/en/ethical-ai-guidelines-binding-commitment-or-simply-window-dressing/

[2] More information on this can be found at https://www.industry.gov.au/data-and-publications/building-australias-artificial-intelligence-capability/ai-ethics-framework/ai-ethics-principles

02

The difference between ensuring ethical and fair AI, and ensuring good decision provenance

Many entities that are implementing automation into decision-making chains have a limited understanding of how to integrate:

- Evaluation of the appropriateness and reliability of automation applications.
- Evaluation of decision-making end-to-end, from inputs to machine-generated outputs (and to what extent those outputs are used to assist people to effect outcomes, or machines are permitted to actuate outcomes).

Frameworks and methodologies for evaluating the end-to-end chain of decision-making – the chain of decision provenance – must be closely aligned with the frameworks and methodologies that focus upon the automation application component or stage within that chain.

Unless there is this close alignment between the automation-specific and end-to-end decision chains, the narrower (but critically important) focus upon the automation application component or stage within a decision chain may not prevent that organisation from delivering unsafe or otherwise unreliable or unacceptable outcomes. Decision provenance needs to be looked at holistically, rather than siloing the automation component and assuming it 'just works' in the context of the larger decision chain.

ACHIEVING ALIGNMENT OF THE AUTOMATION-SPECIFIC AND END-TO-END DECISION CHAINS

Deep integration of automation applications into critical decision paths within many entities is relatively novel. Frameworks and methodologies are, by and large, in the early stages of development.

Development of best-practice guidelines and technical standards for AI systems is proceeding apace, but has principally focussed upon the AI system itself, not the broader organisations into which the AI systems are deployed. Unfortunately, it will take some years to settle the component-specific frameworks for evaluation and governance of AI and other automation applications and then to combine those frameworks with organisation-wide, end-to-end decision chain evaluation and governance.

Notably, alignment of decision chain evaluation and governance requires effective cooperation and a good understanding between technical personnel and other managerial personnel.

Many non-technical executives, including risk management specialists, are still developing the necessary skills and competencies to ensure that alignment. The data inputs, analytics capabilities and technologies driving automation applications are moving faster than the skills and competencies of many senior managers.

Data scientists and other information professionals will need to empower cooperation between technical personnel and other managerial personnel to effect good end-to-end decision chain evaluation and governance. Risk professionals – lawyers, privacy professionals, business risk and data privacy professionals, ethicists and corporate social responsibility specialists – can provide valuable guidance and input.

Ultimately, however, design and specification of good end-to-end decision chain evaluation and governance is a general management function, and must be sponsored and driven as a core organisation-wide activity.

03 Is the inclusion of a 'human in the loop' enough?

Good design of decision provenance systems often requires senior managers within organisations to create guardrails and multiple tracks for decision-making in order to ensure that automated applications are only used when their outputs are reliable.

For good decision-making, there needs to be:

- A culture of caring about societal impacts, not merely addressing automation applications and their use as a compliance issue to be addressed by compliance personnel.
- An understanding of the issues associated with deployment and use of automation applications by those with power and authority to address those issues.
- Foresight and planning to mitigate risks of harm to people or the environment.
- Empowerment of appropriately skilled individuals within an organisation to control how, when and to what extent automation applications are made available, used and relied upon.
- Oversight and review – feedback loops – to evaluate learnings within the organisation and ensure that adverse impacts are promptly identified and lead to re-evaluation of risk mitigations.

Mere inclusion of a 'human in the loop' – within the chain of decision-making between machine output and the application of that output – is not a sufficient control or safeguard unless there is also:

- A clear understanding of respective roles and the relative reliability of people and machines within a chain of decision provenance.
- Appropriate incentives to ensure good outcomes.

Without this understanding and incentives, the 'human in the loop' would be an ineffective buffer against inappropriate use of automation outputs, presenting potential risk for both the organisation and the people and environment with which the organisation interacts.

04 Incentives to *do no harm*

Undesirable outcomes may arise because the organisation using the system does not care whether an automation application causes or contributes to harm to people or the environment.

The starting point for all organisations is ensuring a culture of caring and embodying that culture in consistently reliable processes for determining how, when, why and by whom automation applications are commissioned, deployed, used and relied upon.

An effective culture of caring cannot accept that it is OK to 'move fast and break things' where those actions carry substantial risk of harms to people or the environment.

Technological innovation and the capabilities for outcomes to be affected by automation will inevitably move faster and in ways that the law cannot anticipate and address in advance. The bounds of what is societally acceptable at any time will not be defined by law alone. A culture of caring will supplement the guardrails of what is legally permitted.

An organisation that cares about significant harms to people or the environment should be able to take reasonable steps to determine the nature and extent of risks of harms that may arise from the activities of that organisation.

When risks are unfamiliar, or possible harms not readily apparent, some organisations have started to apply a precautionary principle and undertake further diligence. New models of algorithmic impact assessment, partially derived from now relatively well-developed models for privacy impact assessment, enable organisations to take a structured and comprehensive approach to identification of risks, evaluation of the extent of and possible impact of harms, and to determination of appropriate mitigations.

Whether an organisation cares about causing harm, or does not care, is a function of culture, capabilities, incentives and possible sanctions.

Incentives for the good governance of automation applications varies by industry sector. Particularly powerful and well-understood forms of incentive are exposure to financial penalties or other regulator-imposed sanctions, or to civil liability actions through shareholder or consumer class actions. Legal sanctions are particularly appropriate in relation to the provision or use of products or services when risks of harms are high, or where the organisation does not face sufficiently substantial risks of consumer, citizen or shareholder opprobrium for failure to manage risks.

In many scenarios, incentives may be less legally driven but nonetheless effective. Industry certification schemes and consumer trust mark schemes can provide strong incentives. Fair trade for food products and Energy Ratings for consumer appliances are good examples of market-driven incentives that change the business practices of many suppliers.

05 Regulation and sanctions

Do we need new legal rules and sanctions against the deployment or use of automation applications that pose a significant risk of doing harm to people or the environment?

Whatever an entity's commitment to high standards of social responsibility, digital trust or ethics, that entity needs clear frameworks and methodologies for evaluation of automation applications and of end-to-end decision provenance. These frameworks and methodologies remain works in progress in many industries and organisations.

Until there is better articulation and understanding of frameworks and methodologies for evaluating automation applications, and for evaluating end-to-end decision provenance for decisions where there is automation in the decision chain, organisations cannot be expected to reliably and consistently identify, mitigate and manage residual risks of deployment and use of automation applications.

Formal regulation may be a necessary and proportionate measure to assure good outcomes, at least in those contexts where the deployment of automation poses sufficiently high risk of exposure to harm, or where the level of uncertainty as to assessment and mitigation of risks or harms is unacceptably high. Interim regulation may be appropriate for applications in particular industry sectors or particular use settings where the magnitude of risk of harm is such that a precautionary principle should be applied.

However, overly pre-emptive and restrictive regulation would risk delaying or denying Australian society the benefits of timely availability of automation applications. This is a hard problem to solve, and we can look at some of the global regulatory efforts to get a sense of how to approach this as a nation.

06 Global regulatory efforts

Some other jurisdictions are already considering both new legal liability regimes and new regulatory sanctions as incentives.

In November 2020, the Office of the Privacy Commissioner of Canada (OPC) published its recommendations for a regulatory framework for AI. Pointing out that the use of AI requiring personal information can have serious privacy implications, the OPC made several recommendations, including:

- A requirement for those who develop such systems to ensure that privacy is protected in the design of AI systems.
- A right for individuals to obtain an explanation, in understandable terms, to help them understand decisions made about them by an AI system, including a requirement that such explanations must be based on accurate information and not be discriminatory or biased.
- A right for individuals to contest decisions resulting from automated decision-making.
- A right for the regulator to require evidence of the above.
- A right for the regulator to impose financial penalties on organisations that fail to abide by this regulatory framework.

In contrast to the approach adopted in the European Union's General Data Protection Regulation, the proposed new rights to explanation and contestation would not be limited solely to automated decisions, but would also cover cases where an AI system assists a human decision-maker.

Should these recommendations be adopted in Canada, it would also become necessary for organisations to consider how to explain the decision-making mechanisms, which runs into the explainability problem of complex AI systems (where the 'black box' decision-making process becomes so complex that it can't be explained in human-understandable terms).

A legislated requirement for explainability may lead to concerns that organisations will be forced to disclose proprietary and business confidential algorithms, systems or processes. The OPC noted that "while trade secrets may require organizations to be careful with the explanations they provide, some form of meaningful explanation should always be possible without compromising intellectual property".

The European Commission is also proposing new legislative rules aimed at promoting "excellence and trust in the field of Artificial Intelligence". The declared purpose is "to lay down a balanced and proportionate regulatory approach between the minimal requirements to address the risks and problems linked to AI, without unduly constraining or hindering technological development or otherwise disproportionately increasing the cost of placing AI solutions on the market".

These rules, if enacted, would apply to "both public and private actors inside and outside the EU as long as the AI system is placed on the Union market or its use affects people located in the EU".

The proposed EU regulations define "an AI system" as "software that is developed with one or more of machine learning techniques, logic- and knowledge-based techniques, statistical methods, Bayesian estimations, or search and optimisation methods, and that can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with".

This expansive definition would capture many automation applications that are not 'true' AI or ML and therefore impact a broad range of data-driven businesses.

The proposed new EU legislation comprises:

- Harmonised rules for the use of AI systems impacting people in EU member countries (regardless of where the AI system is situated).
- Prohibitions of certain AI practices that are blacklisted as particularly harmful.
- Specific requirements for high-risk AI systems and obligations for operators of such systems.
- Harmonised transparency rules for AI systems that are intended to interact with individuals, such as emotion recognition systems, biometric categorisation systems, and AI systems used to generate or manipulate image, audio or video content.
- Rules on monitoring and surveillance of the market for provision and use of AI systems.

This framework follows a risk-based approach and differentiates the uses of AI according to whether they create an unacceptable risk, a high risk or a low risk. In the EU's view, risk is unacceptable if it poses a clear threat to people's security and fundamental rights and is prohibited for this reason.

The European Commission identified examples of unacceptable risk as uses of AI that manipulate human behaviour and systems that allow social-credit scoring. For example, this legal framework would prohibit – blacklist – deployment and use affecting EU residents (regardless of the location of the system) of an AI system similar to China's social credit scoring.

High-risk AI systems – including existing systems modified after the effective start date – would need to comply with the rules. Listed high-risk AI systems include systems related to:

- Critical infrastructure (such as road traffic and water supply).
- Educational or vocational training (e.g. the use of AI systems to score tests and exams).
- Biometric identification systems.
- Law enforcement surveillance systems.
- Product safety components of products (e.g. robot-assisted surgery).
- Selection of employees (e.g. resume-sorting software).

AI systems that fall into the high-risk category would be subject to strict requirements.

Among these requirements are:

- Adoption of adequate risk assessment, risk management and quality management systems.
- Compliance with certain data management requirements (related to ensuring data quality and representativeness).
- Creating extensive technical documentation (including demonstrating third party review and compliance).
- Maintaining certain records during system use (including logs of incidences).
- Conducting conformity assessments before use (demonstrating compliance with applicable existing EU laws).
- Third-party compliance checks.

Organisations will also have to design their AI systems to meet certain accuracy, robustness, transparency and cybersecurity standards; enable their outputs to be interpretable by users; and ensure human-in-the-loop capabilities during use.

Like the GDPR, the proposed AI regulations provide for significant penalties for rule violations, including administrative fines. For example, deploying a prohibited AI system or not complying with certain data requirements could result in fines up to EUR 30 million or 6% of a company's "turnover" (revenue). Other specific violations could result in fines of up to EUR 20 million or 4% of a company's revenue (lesser violations would be limited to EUR 10 million or 2% of annual turnover).

Even low-risk AI systems would need to comply with transparency obligations. Users would still need to be aware that they were interacting with a machine. For example, in the case of a 'deepfake', where a person's images and videos are manipulated to look like someone else, it would need to be declared that the image or video content has been manipulated. The European Commission draft does not propose to regulate AI systems that pose little or no risk to European citizens, such as AI used in video games.

The European Commission also proposes to support innovation through so-called 'AI regulatory sandboxes' for non-high-risk AI systems. This would provide a legal environment that facilitates the development and testing of innovative AI systems.

In the United States, the Federal Trade Commission (FTC) has offered business guidance on AI and algorithms, and how companies can manage the consumer protection risks of AI and algorithms. The FTC emphasises that the use of AI tools should be transparent, explainable, fair and empirically sound while fostering accountability. The FTC notes that the use of AI technology to make predictions, recommendations or decisions has great potential to improve welfare and productivity. However, it also presents risks, such as the potential for unfair or discriminatory outcomes or the perpetuation of existing socioeconomic disparities.

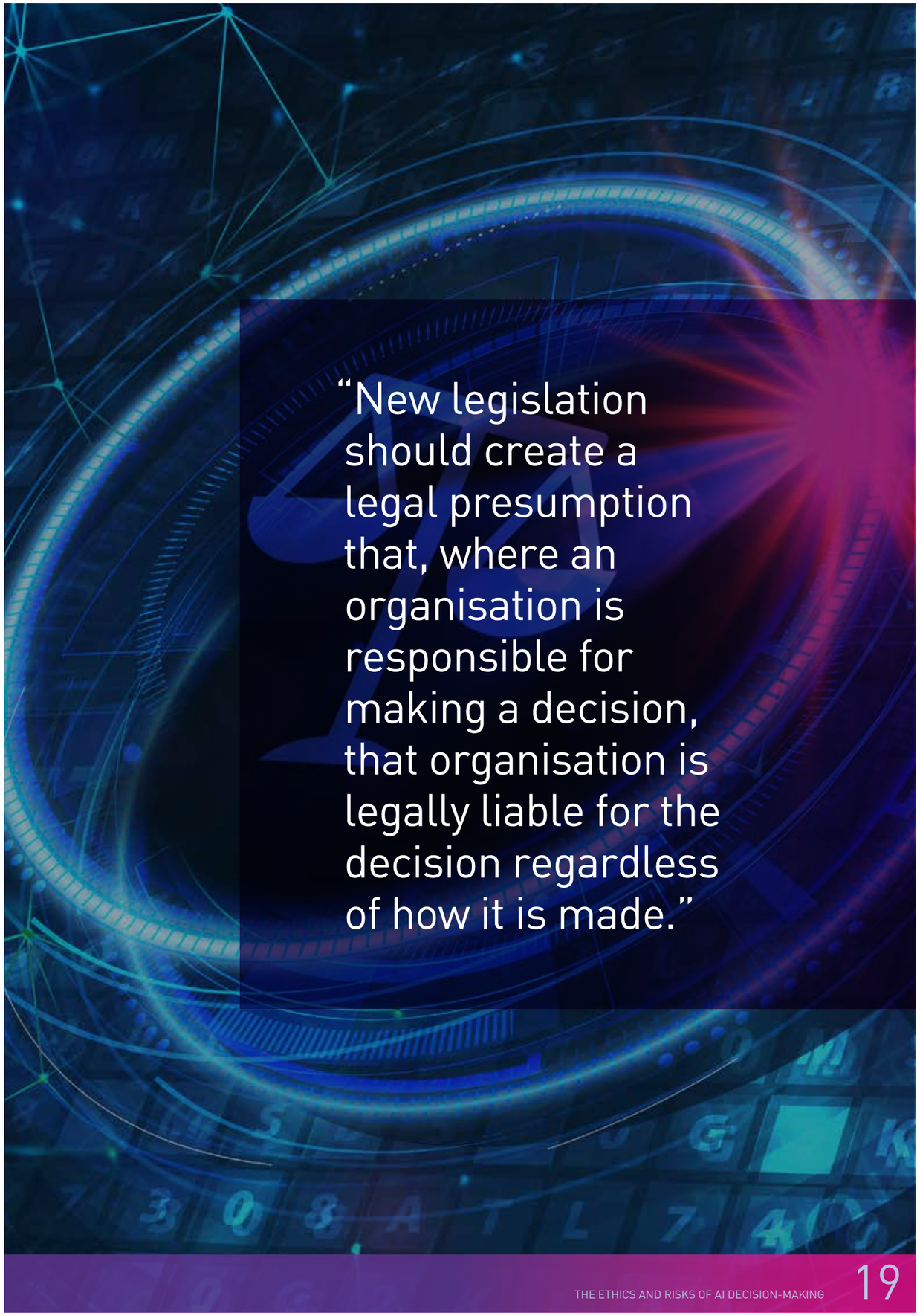
07

The Australian Human Rights Commission report

The Australian Human Rights Commission recently completed a multiyear review of human rights and implications of new technology. The *Human Rights and Technology Final Report* was released in late May 2021. The Commission's interim outputs included a detailed study of how algorithmic bias may contravene legally protected human rights and increase societal inequality.

The Commission's final report includes the recommendations:

- That the Australian Government should use published AI Ethics Principles to encourage corporations and other non-government bodies to undertake a human rights impact assessment before using an AI-informed decision-making system.
- That the government should establish an AI Safety Commissioner and that this Commissioner should issue guidance to the private and public sectors on how to undertake human rights impact assessments.
- That new legislation should require that any affected individual be notified when a corporation or other legal person materially uses AI in a decision-making process that affects the legal, or similarly significant, rights of the individual.
- That new legislation should create a legal presumption that, where an organisation is responsible for making a decision, that organisation is legally liable for the decision regardless of how it is made, including where the decision is automated or is made using AI.
- That new legislation should confirm that where a court or other authority has power order the production of information or other material from an organisation, it must comply with this order even where use of an automation application makes it difficult to comply with the order; and if the person fails to comply with the order because of the technology the person uses, the court or other authority may draw an adverse inference about the decision-making process or other related matters.



08

Building new capabilities for good governance of automation applications

Along with the discussion about the need for new regulations addressing the use of automation applications, there is continuing debate among AI technical and governance experts as to the appropriate range of practical, operational controls, safeguards and guiderails to ensure that outputs from automation applications will be safe, reliable and socially beneficial for use in particular contexts.

There is a need to promote dialogue to build a broad consensus as to how organisations – large and small, and across all sectors of the Australian economy – demonstrate that they are trustworthy in how they make decisions about the design, deployment and use of automation applications.

Demonstrable trustworthiness may require new laws and further regulation to address particular contexts and deployment scenarios.

It will require that organisations be accountable for evaluating the risks of harm and mitigating these risks before automation applications cause harm. Accountability requires clear allocation of responsibilities for governance of automation applications to specified individuals within those organisations.

Good governance of automation applications should become an enduring source of differentiation and competitive advantage for those entities that can demonstrate systemically embedded and reliable adoption of good governance.

REQUIREMENTS FOR GOOD GOVERNANCE OF AI

Whenever legal obligations or evolving standards as to good behaviour by organisations are expected to be given effect by those organisations, three requirements need to be addressed.

First, senior management and boards should understand what is expected of the entity and accordingly of their management and oversight, particularly around strategy and risks.

Many executives and boards of directors now recognise that consideration of social responsibility, protection of enterprise value and maintenance of digital trust creates an imperative for an organisation to reliably conduct a broader and more systematic evaluation of the impacts of an organisation's activities upon individuals and society. However, often executives and boards don't yet know how to specify when an impact assessment should be required, who to empower to undertake that assessment, or how to ensure that impacts are appropriately evaluated and risks reasonably mitigated.

Second, there must be designation and empowerment of appropriately skilled individuals who can give effect to those obligations and standards. Responsibility, skills, incentives, escalation criteria and reporting lines should be properly aligned. Controls, safeguards and properly approved processes and procedures should be reliably embedded within an entity.

Third, those individuals that are given responsibilities for identifying and mitigating risks of adverse impacts from deployment and use of automation applications should be empowered with methodologies and tools that enable those individuals to reliably and verifiably fulfil those responsibilities.

These requirements are common across organisations, regardless of industry sector, size or level of maturity in information technology. Giving effect to these organisational requirements must, however, differ from entity to entity.

Large organisations electing to use automation-enabled decision-making, or to make available automation application-enabled products or services, may have a wide range of capabilities and available resources and may be able to adapt existing functions and processes. A startup, on the other hand, would need to implement these organisational requirements in a quite different way. However, in all cases demonstrable trustworthiness requires each of these organisational requirements to be met.

Implementation of good governance of automation applications often requires upskilling of particular individuals within the organisation, new allocations of responsibilities and reporting lines, and other changes to technology and governance.

These challenges create substantial opportunities for organisations to differentiate themselves from competitors. Some organisations can be expected to be only reactive to public relations crises, to only address existing legal obligations, or be resistant to changes within the organisation. Many organisations would not fully understand the scale of change required, or the imperative to effect change, but in the long term failure to address these risks would corrode trust in government, organisations and data analytics applications.

09

The role of data scientists and other information professionals

An important theme of this paper is the need for an organisation using an automation application to ensure that its frameworks and methodologies for evaluating the end-to-end chain of decision-making – the chain of decision provenance – are closely aligned with the frameworks and methodologies that focus upon the automation application component or stage within that chain.

Some organisations develop and deploy automation applications themselves. Within these entities, there is usually separation between:

- The team that designs an automation application.
- The team that specifies the deployment and operational environment within which the application is used.
- The team that uses the outputs from that application to determine outcomes affecting people or the environment.

Separations such as this can lead to teams making incorrect or otherwise inappropriate allocations of responsibility for decisions; in particular, for ensuring that the right automation application is only used by the right person in the right way for the right job, and in the right workplace and with the right controls and safeguards.

Often automation applications are provided as an ongoing, as-a-service, business-to-business service. Involvement of multiple entities in a chain of decision provenance creates further challenges for good governance.

Errors or mishaps in the use of automation applications can arise in a variety of ways, including:

- 'Pure' human error – failure to properly exercise judgement, and in particular using an application as a determinative instrument rather than as a tool or aid to good human decision-making.
- Excessive or inappropriate reliance upon an application service that is only reliable for use in some particular settings, or only for some classes of decisions.
- Inherent (product) defects in the application itself.

When it comes to data scientists and other information professionals, they often operate on both the supply side – designing, specifying, testing and supplying automation applications – as well as the demand side – specifying the environment in which automation applications are used, the teams that use an application, and evaluation of reliability of suitability for reliance of outputs from use of automation applications.

Straddling both side of the divide, these professionals will likely be the key to the ethical and practical governance of automated decision-making processes, and will often be required to assess whether data inputs are appropriate and reasonable for algorithmic methods to produce reliable outputs, and may need to take the responsibility of flagging inappropriate use of outputs and data.

As we develop more complex AI systems, it's critical that the new generation of IT professionals understand their responsibilities in this regard, and are prepared and capable of guiding their organisations through the transition.

AN ACS LIVING GUIDE

IT professionals need tools and methodologies to help them guide and manage the transition to AI decision-making. To that end, ACS has started work on developing a living guide, to be regularly updated, that will provide:

- Clear definitions of what good assessment and governance of AI looks like according to current best practice.
- Models as to who, within an organisation, should be responsible for which parts of ensuring that the principles and obligations are met.

Our guide will align with:

- Risk management frameworks already in common use (such as the Five Safes).
- Commonly used project management methodologies (such as PRINCE2 and Agile).
- Developing Australian and international standards.

Where such alignment is not possible, we will identify reasons for a change in or expansion of the existing risk management framework or project methodology.

The living guide will provide checklists or tables of questions necessary to ensure that ethical systems are deployed with due regard to social impact; that there's accountability and clarity as to who is responsible for what; and that there's transparency in the deployment, use of outputs and human judgements being applied in the use of the AI.

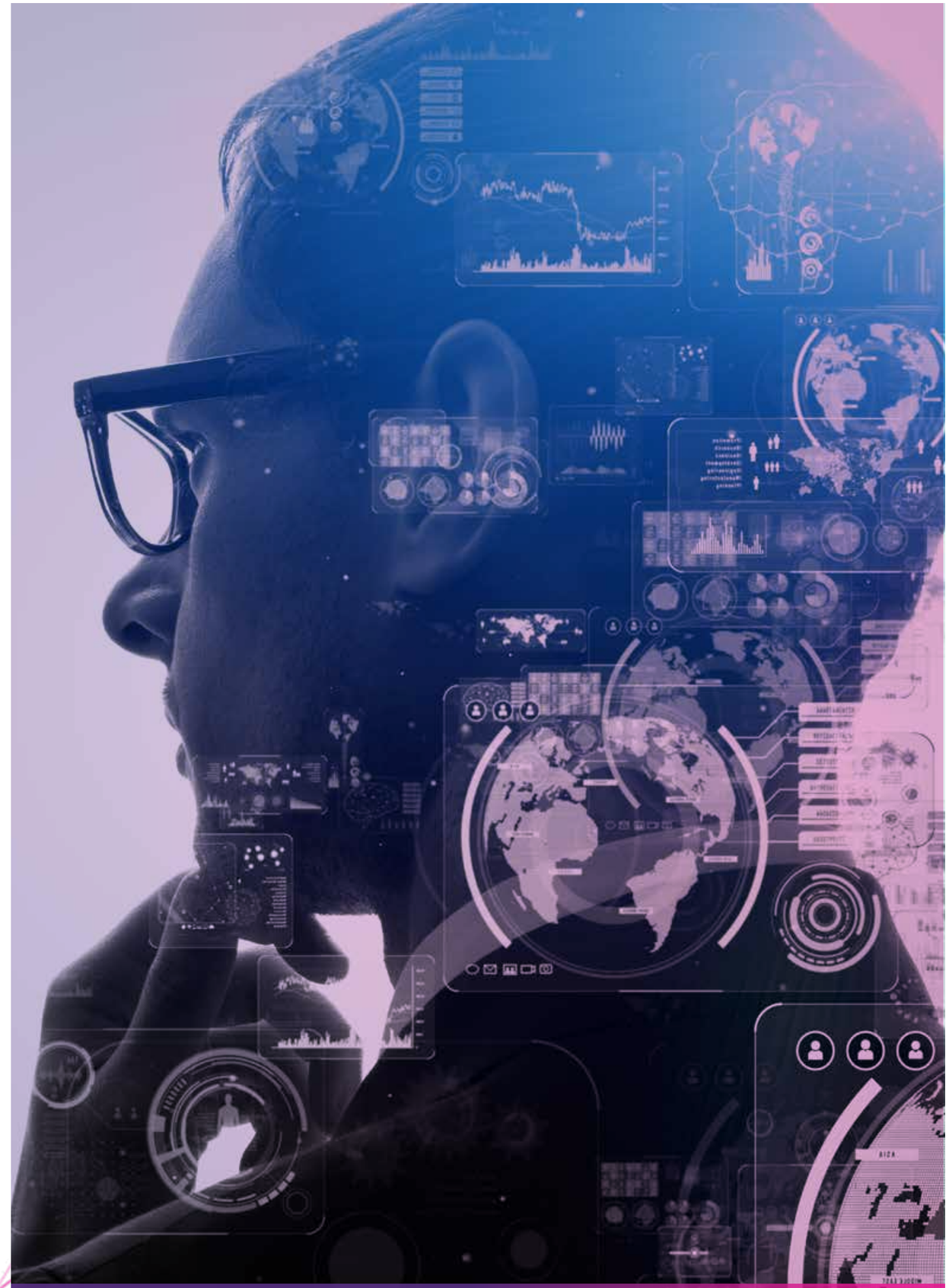
10 Where to next?

Technological innovation in automation will move faster and in ways that the law cannot anticipate and address in advance. IT professionals should not expect the law to be a complete and reliable guide as to how and when use of AI automation are acceptable to Australian society.

That being said, the benefits of AI, ML and other automation-assisted decision-making need not come at a cost. Risks are manageable, but will not be managed without foresight, a culture of caring and good governance, and suitable tools and methodologies for assurance of safety and reliability of decision-making processes in which automation applications are a component.

Existing governance processes for structuring and oversight of decision-making within an organisation will often require a significant overhaul to address new deployments and uses of automation applications. This requires careful consideration of the interaction of people, technology, data and processes.

Through this, ACS will provide IT professionals with the tools and methodologies needed to assist them in this transition.



Further reading

DETERMINING WHETHER AN AUTOMATION APPLICATION IS 'GOOD AI': PRINCIPLES AND FRAMEWORKS

Australian Human Rights Commission, *Human Rights and Technology Final Report*, May 2021
https://tech.humanrights.gov.au/downloads

Australian Government Department of Industry, Science, Energy and Resources, *AI Ethics Principles*
https://www.industry.gov.au/data-and-publications/building-australias-artificial-intelligence-capability/ai-ethics-framework/ai-ethics-principles

Algorithm Watch, *AI Ethics Guidelines Global Inventory*, as updated at
https://algorithmwatch.org/en/project/ai-ethics-guidelines-global-inventory/

Algorithm Watch, *Ethical AI guidelines: Binding commitment or simply window dressing?*
https://algorithmwatch.org/en/ethical-ai-guidelines-binding-commitment-or-simply-window-dressing/

OECD, *AI Principles*, May 2019

Monetary Authority of Singapore, *Principles to Promote Fairness, Ethics, Accountability and Transparency (FEAT) in the Use of Artificial Intelligence and Data Analytics in Singapore's Financial Sector*

Singapore Academy of Law, Law Reform Committee, *Applying Ethical Principles for Artificial Intelligence in Regulatory Reform*, July 2020

European Parliamentary Research Service of the European Parliament, *The ethics of artificial intelligence: Issues and initiatives*, March 2020
https://www.europarl.europa.eu/RegData/etudes/STUD/2020/634452/EPRS_STU(2020)634452_EN.pdf

UK Statistics Authority, *Identifying gaps, opportunities and priorities in the applied data ethics guidance landscape, Annex A Landscape Review Table, Key User-focused Data Ethics Activities*, April 2021
https://uksa.statisticsauthority.gov.uk/wp-content/uploads/2021/04/Annex-A-Landscape-Review-Table.pdf

Brent Mittelstadt, *Principles alone cannot guarantee ethical AI*, Nature Machine Intelligence, Vol. 1, November 2019, pp501–507, 2019

UK Government, *Data Ethics Framework*

NSW Government, *AI Ethics Policy*
https://www.digital.nsw.gov.au/policy/artificial-intelligence-ai/ai-ethics-policy

World Economic Forum, *Model Artificial Intelligence Governance Framework and Assessment Guide*
https://www.weforum.org/projects/model-ai-governance-framework

Rohit Satish and Tanay Mahindru (NITI Aayog), *Towards Responsible AI for All, Approach Document for India, Part 1: Principles for Responsible AI*, February 2021
http://www.niti.gov.in/sites/default/files/2021-02/Responsible-AI-22022021.pdf

Republic of Singapore, Infocomm Media Development Authority, *Model AI Governance Framework, Second Edition*, 2020
https://www.pdpc.gov.sg/-/media/files/pdpc/pdf-files/resource-for-organisation/ai/sgmodelaigovframework2.pdf

World Economic Forum and Info-communications Media Development Authority of Singapore, *Companion to the Model AI Governance Framework – Implementation and Self-Assessment Guide for Organizations*, January 2020
https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Resource-for-Organisation/AI/SGIsago.pdf

Singapore Computer Society (SCS), *AI Ethics & Governance Body of Knowledge (BoK)*
https://www.scs.org.sg/ai-ethics-bok

Montreal AI Ethics Institute, *The State of AI Ethics*, October 2020

Jessica Field and others, *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI*, Berkman Klein Center for Internet & Society Research Publication Series, 2020

DETERMINING WHETHER AN AUTOMATION APPLICATION IS 'GOOD AI': CASE STUDIES

Republic of Singapore, Infocomm Media Development Authority, *Compendium of Use Cases Volume 1: Compendium of Use Cases: Practical illustrations of Model AI Governance Framework*
https://www.pdpc.gov.sg/-/media/files/pdpc/pdf-files/resource-for-organisation/ai/sgaigovusecases.pdf

Republic of Singapore, Infocomm Media Development Authority, *Compendium of Use Cases Volume 2: Practical illustrations of Model AI Governance Framework*
https://file.go.gov.sg/ai-gov-use-cases-2.pdf

Monetary Authority of Singapore, *Veritas Document 1, FEAT Fairness Principles Assessment Methodology*, December 2020

Monetary Authority of Singapore, *Veritas Document 2, FEAT Fairness Principles Assessment Case Studies*, December 2020

Centre for Data Ethics and Innovation, *AI Barometer Report*, June 2020

Partnership on AI, *Report on Algorithmic Risk Assessment Tools in the U.S. Criminal Justice System*
https://www.partnershiponai.org/report-on-machine-learning-in-risk-assessment-tools-in-the-u-s-criminal-justice-system/

Georgina Ibarra, David M. Douglas and Meena Tharmarajah, *Machine Learning and Responsibility in Criminal Investigation*, October 2020

ACHIEVING ALIGNMENT OF AUTOMATION-SPECIFIC AND END-TO-END DECISION CHAINS, EVALUATION AND GOVERNANCE

Inioluwa Deborah Raj and others, *Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing*, January 2020

Juan Aristi Baquero and others, *Derisking AI by design: How to build risk management into AI development*, McKinsey Analytics, August 2020

Reuben Binns, *Analogies and disanalogies between machine-driven and human-driven legal judgement*, Journal of Cross-Disciplinary Research in Computational Law, 2020

AI ASSURANCE AND STANDARDS

Centre for Data Ethics and Innovation Blog, *The need for effective AI assurance*
https://cdei.blog.gov.uk/2021/04/15/the-need-for-effective-ai-assurance/

Centre for Data Ethics and Innovation Blog, *User needs for AI assurance*
https://cdei.blog.gov.uk/2021/04/16/user-needs-for-ai-assurance/

Centre for Data Ethics and Innovation Blog, *Types of assurance in AI and the role of standards*, 17 April 2021
https://cdei.blog.gov.uk/2021/04/17/134/

Further reading cont.

Standards Australia, *An Artificial Intelligence Standards Roadmap: Making Australia's Voice Heard*, Final Report, 2020
https://www.standards.org.au/getmedia/ede81912-55a2-4d8e-849f-9844993c3b9d/1515-An-Artificial-Intelligence-Standards-Roadmap12-02-2020.pdf.aspx

ISO, *Standards by ISO/IEC JTC 1/SC 42: Artificial intelligence*
https://www.iso.org/committee/6794475/x/catalogue/

ISO/IEC TR 24028:2020, *Information technology - Artificial intelligence - Overview of trustworthiness in artificial intelligence: Technical Report*, 2020

Peter Cihon, *Standards for AI Governance: International Standards to Enable Global Coordination in AI Research & Development: Technical Report*, 2019

ALGORITHMIC BIAS

Australian Human Rights Commission, *Using artificial intelligence to make decisions: Addressing the problem of algorithmic bias*, Technical Paper, 2020

UK Information Commissioner's Office and Alan Turing Institute, *Project explain: Explaining decisions made with AI*, May 2020

Centre for Data Ethics and Innovation, *Review into bias in algorithmic decision-making*, November 2020

Solon Barocas, Moritz Hardt and Arvind Narayanan, *Fairness in Machine Learning: Limitations and Opportunities*

BCS, The Chartered Institute for IT, *The Exam Question: How Do We Make Algorithms Do The Right Thing?*, 2020

Rebecca Williams, *Rethinking deference for algorithmic decision-making*

EXPLAINABILITY

P. Jonathon Phillips and others, *Four Principles of Explainable Artificial Intelligence*, Draft NISTIR 8312, August 2020

Ada Lovelace Institute and DataKind UK, *Examining the Black Box: Tools for assessing algorithmic systems: Identifying common language for algorithm audits and impact assessments*, April 2020

Georgina Ibarra and Meena Tharmarajah, *Designing trustworthy machine learning systems*, November 2020
https://algorithm.data61.csiro.au/designing-trustworthy-machine-learning-systems/

INCENTIVES, REGULATION AND SANCTIONS

Daniel Montoya and Alice Rummery, *The use of artificial intelligence by government: parliamentary and legal issues*, NSW Parliamentary Research Service e-brief 02/2020, September 2020

Office of the Privacy Commissioner of Canada, *A Regulatory Framework for AI: Recommendations for PIPEDA Reform*, November 2020

Commonwealth Ombudsman, *Services Australia's Income Compliance Program: A Report About Services Australia's Implementation Of Changes To The Program In 2019 And 2020*, April 2021

Commonwealth Ombudsman, *Centrelink's Automated Debt Raising and Recovery System: Implementation Report*, April 2019

European Commission, draft *Regulation On A European Approach For Artificial Intelligence*, April 2021

European Commission, *White Paper on Artificial Intelligence – A European approach to excellence and trust*, COM(2020) 65 final, 2020

European Parliament resolution of 20 October 2020 with recommendations to the Commission on a framework of ethical aspects of artificial intelligence, robotics and related technologies, 2020/2012(INL)

Christopher S. Yoo and Alicia Lai, *Regulation of Algorithmic Tools in the United States*, Journal of Law & Economic Regulation, Vol. 13 No. 2, pp7–22, 2020

Gary Marchant, Lucille Tournas and Carlos Ignacio Gutierrez, *Governing Emerging Technologies Through Soft Law: Lessons for Artificial Intelligence – An Introduction*, Jurimetrics, Vol. 61 No. 1, pp1–18, 2020

Sandra Wachter and Brent Mittelstadt, *A Right To Reasonable Inferences: Re-Thinking Data Protection Law In The Age Of Big Data and AI*, Columbia Business Law Review, Vol. 2019 No. 2, pp494–620, 2019

BUILDING NEW CAPABILITIES FOR GOOD GOVERNANCE OF AUTOMATION APPLICATIONS

Information Commissioner's Office (UK), *Guidance on the AI auditing framework: Draft guidance for consultation*, February 2020
https://ico.org.uk/media/about-the-ico/consultations/2617219/guidance-on-the-ai-auditing-framework-draft-for-consultation.pdf

Information Commissioner's Office (UK), *Toolkit for organisations considering using data analytics*, February 2021
https://ico.org.uk/for-organisations/toolkit-for-organisations-considering-using-data-analytics/

Information Commissioner's Office (UK), *How do we ensure individual rights in our AI systems?*
https://ico.org.uk/for-organisations/guide-to-data-protection/key-data-protection-themes/guidance-on-ai-and-data-protection/how-do-we-ensure-individual-rights-in-our-ai-systems/

Ghazi Ahamat and the ors, *Types of assurance in AI and the role of standards*, Centre for Data Ethics and Innovation Blog, 17 April 2021
https://cdei.blog.gov.uk/2021/04/17/134/

Jatinder Singh and others, *Decision Provenance: Harnessing data flow for accountable systems*, IEEE Access, Vol. 7, pp6562–6574, 2019

NSW Government, *AI User Guide*
https://www.digital.nsw.gov.au/policy/artificial-intelligence-ai/user-guide

Information Accountability Foundation, *The Movement Towards Demonstrable Accountability – Why It Matters*, December 2020

High-Level Expert Group on Artificial Intelligence of the European Commission, *Assessment List for Trustworthy Artificial Intelligence*, July 2020

Office for Statistics Regulation (UK), *The way forward: Reproducible Analytical Pipelines, Overcoming barriers to adoption*, March 2021 https://osr.statisticsauthority.gov.uk/publication/reproducible-analytical-pipelines-overcoming-barriers-to-adoption/

Lisa R. Lifshitz and Cameron McMaster, *Legal And Ethics Checklist For AI Systems*, The Scitech Lawyer, Fall 2020, pp28–34, 2020

World Economic Forum, *Guidelines for AI Procurement*, White Paper, September 2019

ACS

ABN: 160 325 931
International Tower One
Level 27, 100 Barangaroo Avenue
Sydney NSW 2000

P: 02 9299 3666
F: 02 9299 3997

About ACS

ACS is the professional association for Australia's Information and Communication Technology (ICT) sector. Over 48,000 ACS members work in business, education, government and the community. The Society exists to create the environment and provide the opportunities for members and partners to succeed. ACS strives for ICT professionals to be recognised as drivers of innovation in our society, relevant across all sectors, and to promote the formulation of effective policies on ICT and related matters.

Visit **acs.org.au** for more information.