Australia's National
Science Agency

CSIRO submission to Supporting Responsible AI Discussion Paper

CSIRO Submission 23/828

July 2023

Main Submission Authors:

Dr Liming Zhu

Aurélie Jacquet

Dr Qinghua Lu

Enquiries should be addressed to:

E governmentrelations@csiro.au

Contents

Executive Summary

In this submission, CSIRO addresses selected questions in the Safe and Responsible Artificial Intelligence (AI) in Australia discussion paper that relate to CSIRO's scientific and technological expertise. CSIRO proposes seven non-regulatory initiatives designed to increase the adoption of responsible AI practices, thus providing a competitive advantage for the industry and positioning Australia as a world leader in responsible AI.

- **Initiative 1:** Develop industry best practices, playbooks, guidelines, and case studies for Australia's priority industry sectors, especially targeting small and medium enterprises (SMEs), while considering Australia's unique context.
- **Initiative 2:** Develop trustworthiness metrics, measurement, testing, evaluation, verification, and validation (TEVV) methods and guidelines along with associated tools and products, including the seeding of a world-leading responsible AI tool industry in Australia.
- **Initiative 3:** Setup up programs to encourage and incentivise industry and government to develop new validated best practices and share them within the Australian industry and globally.
- **Initiative 4:** Set up a national sandbox to explore and experiment with responsible AI approaches in a safe environment.
- **Initiative 5:** Set up connected responsible AI awareness and training programs.
- **Initiative 6:** Set up a national responsible AI technology program to inform responsible AI policy, regulation, and international standards.
- **Initiative 7:** Identify responsible AI approaches and edge cases that can benefit all Australians.

Finally, there are several essential points to consider across all initiatives:

- Emphasise AI governance at the system level, not just the model level.
- Pay particular attention to the intersection of AI with other vital and emerging technologies such as cybersecurity, quantum systems, blockchain, and robotics.
- Concentrate on the empirical understanding and experimentation of AI uses and technologies.
- Adopt a supply chain perspective.

CSIRO would welcome the opportunity to discuss these matters in more depth with the Department of Industry, Science and Resources.

Introduction

CSIRO welcomes the opportunity to provide a response to the Safe and Responsible AI in Australia discussion paper.

As Australia's national science agency, CSIRO is at the centre of solving Australia’s greatest challenges through innovative science and technology. CSIRO contributes to inclusive, ethical, safe and secure AI adoption in Australia through research in areas of advanced AI technologies, AI engineering, responsible AI and innovation, human-AI collaborative intelligence, AI-enabled Assistive Technologies, regulation technology, privacy and cybersecurity of AI systems.

CSIRO hosts the National AI Centre to helps grow an AI industry in Australia and helps Australian industry responsibly adopt AI in strategically important sectors, such as manufacturing, energy, agriculture, critical minerals, digital services and defence.

CSIRO welcomes the opportunity to further discuss this submission – and collaboration opportunities – with the Department of Industry, Science and Resources.

Responses to Discussion Paper Questions

1.1 Definitions

Q1: Do you agree with the definitions in this discussion paper? If not, what definitions do you prefer and why?

The definitions used in the discussion paper provide a reasonable stipulative explanation for the purposes of the responsible AI discussion. However, depending on the context in which the definition is used, a tailored definition emphasising operational ease is crucial.

Defining a concept or term can be approached in multiple ways, each having its own advantages and disadvantages. Given that the main goal of responsible AI governance is to operationalise the high-level principles, standards, frameworks, and regulations efficiently and effectively in specific contexts, it is important to focus on the operational practicality of the definition. It should be straightforward to discern whether a particular case falls within the definition. Because the goal is to operationalise AI governance, international policy initiatives such as the US National Institute of Standards and Technology (NIST), OECD and ISO's use a broad definition of AI systems:

- **OECD definition:** An AI system is a machine-based system that is capable of influencing the environment by producing an output (predictions, recommendations or decisions) for a given set of objectives (we note that the OECD is making some changes to its definition of AI as presented and discussed at the latest OECD meeting in Paris in April 2023).
- **ISO definition:** AI system is an engineered system that generates outputs such as content, forecasts, recommendations or decisions for a given set of human-defined objectives.
- **NIST definition:** The AI RMF refers to an AI system as an engineered or machine-based system that can, for a given set of objectives, generate outputs such as predictions, recommendations, or decisions influencing real or virtual environments. AI systems are designed to operate with varying levels of autonomy (Adapted from: OECD Recommendation on AI:2019; ISO/IEC 22989:2022).
- **European Union AI Act definition (in the latest compromise text 16/05/2023):** For the purpose of this Regulation, the following definitions apply:
    - 'artificial intelligence system' (AI system) means a machine-based system that is designed to operate with varying levels of autonomy and that can, for explicit or implicit objectives, generate outputs such as predictions, recommendations, or decisions that influence physical or virtual environments.

In addition, the context of where the definitions are applied can also affect their nature. For instance, a definition used in legal documents, standards, policy papers, or technical practices might vary. A legal definition might demand precision regarding inclusiveness/exclusiveness that necessitates the appropriate considerations of exceptions.

In non-regulatory environments, where system-level responsible development involves both AI-enabled components and non-AI components, understandability and flexibility may take precedence over precision and over-inclusiveness. Similarly, in AI research, a definition typically relies on theoretical frameworks and utilises a genus-difference approach to position the concept within a broader context.

Due to the rapidly evolving, complex, and general-purpose nature of AI technologies, many global entities suggest a risk-based approach over others (refer to answers in Q14 and Q15). In such an approach, risk sources are considered to identify if an AI system is to be categorised as low, medium or high risk[1]:

- **Risks related to technical approaches –** how AI is constructed, such as technical approaches like reinforcement learning, supervised/unsupervised learning, data-driven, or symbolic approaches.
- **Risks related to use cases –** the intended applications of AI.
- **Risks related to its general capability –** what AI can accomplish, ranging from intrinsic measures like training data size, computing capacity used, and model size, to general capabilities such as reasoning, common sense, and reading comprehension.

Incorporating risk sources both categorically and with detailed enumeration can effectively help practical operational sense, especially since many governance strategies and technological controls are risk-source related.

We further suggest the following changes to definitions for consideration:

- **Machine Learning:** Instead, consider **Machine Learning Model**. This term would more accurately represent the "patterns derived from training data using machine learning algorithms, which can be applied to new data for prediction or decision-making purposes." It would be more accurate to refer to the entity that performs a specific task as a model instead of AI.
- **First Nations:** Instead, consider 'Aboriginal and Torres Strait Islander' (see p.8). The use of First Nations is an internationally broad term and does not acknowledge the cultural diversity of Aboriginal and Torres Strait Islander peoples.

[1] https://www.europarl.europa.eu/resources/library/media/20230516RES90302/20230516RES90302.pdf

1.2 Potential gaps in approaches

Q2: What potential risks from AI are not covered by Australia's existing regulatory approaches? Do you have suggestions for possible regulatory action to mitigate these risks?

There is limited consideration around the use of AI with vulnerable populations, such as aging and people living with a disability. While the Therapeutic Goods Administration (TGA) includes regulation for medical devices, there are AI applications that are not intended as medical devices but are in fact used by people to support their daily life without proper consideration and understanding of the risks and with limited tools to mitigate the risks. The CSIRO and National Disability Insurance Agency (NDIA) framework for AI-enabled Assistive Technology proposed a framework to mitigate these risks (Silvera et. Al, 2022).

There are also flaws in available data regarding Aboriginal and Torres Strait Islander peoples. Regulation is needed to ensure the positive effects of AI are accessible and experienced by all Australians in healthcare. The risks are not solely due to algorithmic biases (as mentioned in the discussion paper) but also due to the data biases inherited from the data sets an AI model is trained on. Data quality and fairness are pivotal for the success of any AI-based solutions. The mismanagement of data biases poses a potential source of discrimination and injustice. Unfortunately, due to Indigenous Data Paradox, Indigenous healthcare data is commonly misrepresented and negatively biased. Responsible and ethical guidelines reflecting Indigenous Data Sovereignty need to be enshrined into the existing framework (Walter, 2018). To achieve this, we suggest a consultative approach ensuring a health justice approach to AI data governance.

The lack of data is particularly pronounced in genomics and extends to all non-European backgrounds (Mills, 2020). Genomic information increasingly informs healthcare and the lack in actionable markers for non-Caucasian populations threatens to increase the health disparity. For example, a CSIRO study found that 10% of cystic fibrosis variants were missed in Australia's multiethnic population (Shum et al. 2022). Machine Learning (ML) powering automated data annotation and pre-diagnostic advise will be influenced from this disparity in knowledge and data. Data is constantly changing, so do the inputs into AI models, and therefore AI-based systems need continuous evaluation and mechanisms to detect and action on the decline in performance. None of the proposed regulations covers continuous performance monitoring and evaluation.

The policies related to AI are too difficult for industry players to demonstrate conformity with based on internal processes only. Therefore, a third-party auditing of AI-based service/product providers is needed. Auditing by an independent, specialised entity facilitates the compliance process, especially for smaller players. We note that a number of jurisdictions (United Kingdom, European Union, Canada), are considering the certifications of AI systems as an appropriate tool to ensure responsible AI.

Q3: Are there any further non-regulatory initiatives the Australian Government could implement to support responsible AI practices in Australia? Please describe these and their benefits or impacts.

The following non-regulatory initiatives could significantly benefit Australia in adopting responsible AI practices, thus providing a competitive advantage for the industry and positioning Australia as a world leader in responsible AI.

Initiative 1: Develop industry best practices, playbooks, guidelines, and case studies for Australia's priority industry sectors, especially targeting small and medium enterprises (SMEs), while considering Australia's unique context.

Given the recent G7 call for the adoption of international technical standards on AI, the release of both AI international and country/region-initiated standards (ISO, the Institute of Electronic and Electronics Engineers (IEEE), NIST, the European Committee for Standardization (CEN) and the European Committee for Electrotechnical Standardization (CENELEC), and the European Telecommunications Standards Institute, etc.), and the related certification initiatives (such as the certification of high-risk AI systems under the European Union’s AI Act, the upcoming conformity assessment of ISO 42001, and IEEE's CertifAIEd), there is an increasing need to take stock of international initiatives and equip the Australian Industry with sector-specific, SME-friendly guidance based on practical and tangible use cases. This could help to:

- Bridge the gap that currently exists between horizontal international standards certification initiatives, and the vertical practices of industry sectors; and
- Encourage and accelerate the responsible adoption of AI in Australia that is interoperable with internationally recognised best practice.

CSIRO scientists, along with 35 other nations (including the European Union, United States, United Kingdom, India, Japan, Canada), are already playing a key role in shaping the ISO Standards on AI and participating in other international initiatives (such as OECD AI expert groups). CSIRO also collaborates with organisations like the NIST that focus on providing resources to address the post-standard/regulation implementation gap. For instance, CSIRO’s responsible AI pattern catalogue has aided Australian companies from different sectors in enhancing their practices (CSIRO, 2022).

There is a potential to coordinate this sector-specific approach with like-minded countries, allowing each nation to lead in their respective priority sectors, contribute inputs, and share results and relevant implementation practices.

Such coordination could streamline the process of standard conformance, regulation compliance, and other responsible AI adoption procedures, making it more cost-effective, particularly for SMEs. It could also better equip Australian companies to scale up and compete globally, which is especially significant for Australia's priority sectors and use cases.

Initiative 2: Develop trustworthiness metrics, measurement, Testing, Evaluation, Verification, and Validation (TEVV) methods and guidelines along with associated tools and products, including the seeding of a world-leading responsible AI tool industry in Australia.

Regardless of the form of regulation and governance, an increasing need is arising:

- From the individual and community perspective, to understand the trustworthiness and associated risks of digital products/services; and
- From an organisational perspective, to demonstrate the trustworthiness of their digital products/services.

For these reasons, and with the goal of complementing high-level international AI standards and frameworks, there is a growing international interest in developing metrics, measurements, and TEVV methods to evaluate the trustworthiness of AI systems using standardised tools. Many existing TEVV methods are highly manual and therefore unfit for the era of AI, which often demands continuous assessments due to its dynamic nature. The manual effort and dependence on high-level human expertise will particularly hinder SME adoption and competitiveness, as they cannot afford expensive third-party consulting services.

Australia can take the lead on a significant portion of this effort by establishing programs to develop metrics, measurements, and more automated TEVV methods to accelerate and expand AI adoption, particularly for SMEs. This should include considerations for individuals and the community, as they also cannot afford consulting services to mitigate risks. This is particularly important in vulnerable minority/vulnerable populations not often considered in an organisational assessment. These methods and associated tools can also indicate best practices for risk mitigation and assurance enhancement.

For instance, to achieve tangible AI risk measurement, CSIRO is developing a question bank that contains questions and potential metrics from existing AI risk assessment standards and frameworks (Lee 2023, Xia 2023).

CSIRO also collaborates with the NDIA to address the assessment gap and required metrics in AI applications when used by people with a disability and other vulnerable populations (Silvera et. Al, 2022).

With clear metrics/measurements and cost-effective, semi-automated TEVV, the Australian industry can expedite the rollout of quality AI-enabled products and services, showcasing demonstrable measures as a competitive advantage. Stakeholders in the wider community, the public, and regulatory bodies can monitor relevant metrics and examine the measurement/TEVV methods to gain a better understanding of the trustworthiness of these products and services.

In addition, such a program has the opportunity to seed a world-leading responsible AI tool and product industry in Australia.

Initiative 3: Setup up programs to encourage and incentivise industry and government to develop new validated best practices and share them within the Australian industry and globally.

Due to the significant complexity and rapid evolution of AI technologies, it is not always the case that the best practices are already known and simply need adoption. Many best practices, including those specific to certain sectors, contexts, or use cases, are being created through innovations and experiments in responsible AI. These practices need to be regularly updated or reinvented as AI technologies evolve.

While some of these best practices can become commercially sensitive competitive advantages for a company, many others can be shared within the Australian industry and globally to uplift the ecosystem. Companies need incentives to do this, alongside trusted and scientifically rigorous validation of these best practices. For example, CSIRO has been creating and curating a set of reusable responsible AI best practices through a responsible AI pattern catalogue. Some of these best practices are extracted or updated through our collaboration with leading Australian companies practising responsible AI in their specific contexts (Lu et al. 2023).

This incentive to develop new best practices will underpin Australia's international leadership in responsible AI.

Initiative 4: Set up a national sandbox platform to explore and experiment with responsible AI approaches in a safe environment.

As the list of responsible AI best practices and regulatory initiatives grows rapidly, there is a pressing need to provide the industry with a platform where they can experiment and identify the Responsible AI approach that best suits their unique context. This would enable them to create high-quality, innovative AI-enabled products and services. A national sandbox platform could present a unique opportunity to trial innovative strategies for scaling AI responsibly, in collaboration with trusted third parties, thereby potentially positioning Australia as a leader in developing top-tier AI products and services.

Another possible application of the sandbox could involve setting up a government-as-example initiative. This would not only coordinate government implementation of responsible AI but also pioneer best practices that could be utilised by the broader Australian industry.

Initiative 5: Set up connected responsible AI awareness and training programs.

AI risks are often evaluated and managed in isolated units within organisations. For example, AI risks pertaining to privacy, security, Health Safety Environment (HSE), and Environmental, Social, and Governance (ESG) are frequently handled by various teams and functions at different levels. Training and awareness programs conducted by distinct organisations often target these specific units, leading to resource wastage and internal resource competition. On the other hand, many best practices mitigations could reduce multiple risks simultaneously if implemented organisation-wide appropriately.

There lies an opportunity to establish interconnected responsible AI awareness and training programs, aiming to dismantle these silos while delivering relevant, tailored information to satisfy the specific needs of stakeholders.

For instance, CSIRO's responsible AI pattern catalogue was deliberately designed to connect different perspectives and levels (CSIRO, 2022). It aligns practices for both the development process and the AI product with the governance practices at organisational and executive levels. CSIRO's cybersecurity and privacy research groups are working with the responsible AI research group to adopt an integrated approach to AI risks related to privacy and security, including within supply chain contexts. We have also initiated a new project that incorporates an examination of AI risks in the context of ESG[2]. Finally, CSIRO is constructing an integrated question bank intended to enhance comprehension of AI risks, ranging from engineering to boardroom levels.

Demolishing these silos and integrating risk understanding holistically across the organisation will not only reduce costs but also significantly enhance efficiencies and assurance in implementing AI responsibly.

Initiative 6: Set up a national responsible AI technology program to inform responsible AI policy, regulation, and international standards.

Given the complexity and rapid evolution of AI technologies, a strategic and systematic approach is crucial for experimenting and gathering evidence that can better support and inform policy, regulation, and standards.

There is an opportunity to establish a long-term AI technology program strategically designed to inform and support Australia's policy, regulation, and international contribution. The governance body managing such a program should include representation from all relevant parties, including government, industry, human rights, legal, research, and diverse representation of end user groups, including vulnerable populations (e.g., organisations that represent people with a disability). This technology program would not only provide advice based on expertise but could also possess technical resources to experiment with policy/regulatory/standards options and generate evidence to support its advice. This approach is similar to that of the NIST – the "Technology" aspect of NIST allows them to provide inputs strategically and systematically to standards and related policy and regulation questions. Example activities of such a technology program could include elements of Initiatives 1, 2, 3, and 4 mentioned above, with the purpose of providing inputs into policy, regulation, and international standards based on Australian priorities and context.

Some industries may be hesitant to directly share their data, models, practices, and issues with a regulator or responsible government agency. However, this information is crucial for both research and policy/regulation/standard-setting. Using a national, research-centric technology program as a trusted intermediary could help to facilitate the anonymised sharing of sensitive data, models, practices, and issues for responsible AI.

[2] https://www.csiro.au/en/news/all/news/2023/june/alphinity-and-csiro-partnership-media-release

An example of such an effort is the Department of Home Affairs' 6G security program with CSIRO. The program focuses on fundamental technical research into 6G security and its intersection with underlying AI/ML technologies, specifically aiming to strategically inform connected policy, regulation, and standards.

Finally, there are other approaches that could serve as alternatives or complements to risk-based approaches, depending on the aspects of AI governance being considered. Please refer to responses to Q14 and Q17 for more details. The technological program and the sandbox initiative can aid in experimenting with these alternative approaches, providing science and data-driven insights into integrated strategies for AI governance.

Initiative 7 - Identify responsible AI approaches and edge cases that can benefit all Australians.

With AI technology there is an increased incentive and temptation to automate for the majority rather than the edge case (e.g., minority populations):

- As AI learns from data, it becomes easier to automate for an individual or group of individuals that are highly visible in the data. Those that are invisible in the data, are often instead considered as blind spots, e.g., refugees are often missed in a census or survey.
- The cost of training AI systems and the financial return on automation encourages organisations to develop models that operate well for the majority with some adaptation or recourse for the edge cases rather than developing AI systems that are designed for the edge cases, the "data poor" or the "data edge cases".

There is an opportunity to incorporate product design practices and design for the underrepresented populations to promote better innovation that could benefit the broader society, including minority groups[3].

Considering the above, the first step would be to establish a research initiative that advocates for the development of AI systems benefiting all Australians. This initiative could be carried out by a trusted research organisation that works with partners and communities to identify the best strategies and edge cases (such as underrepresented populations) for integrating the values, culture, knowledge, and knowledge acquisition approaches of minority groups. A good example is CSIRO's work in collaboration with the NDIA on the AI-enabled Assistive Technology framework for people living with a disability (Silvera, 2022). This project involved significant engagement with industry, end users and representatives of the disability community. Similarly, projects within the initiative might be led by the respective minority group and tailored specifically to benefit those identified as "data minorities". Examples of such groups could include Aboriginal and Torres Strait Islander communities, the LGBTIQA+ community, communities with culturally and linguistically diverse backgrounds, people with a disability and rural or regional communities.

[3] https://medium.com/dayone-a-new-perspective/how-the-fear-of-edge-cases-kills-ideas-71413a2ff59d

Finally, there are several essential points to consider across all initiatives:

- **Emphasise AI governance at the system level, not just the model level:** The system, along with the products and services, is what ultimately impacts individuals, groups, and communities. Australia has established a leadership role in the field of system-level responsible AI governance, including responsible AI engineering practices and guidelines. This advantage could be harnessed for responsible AI adoption and to seed a responsible AI tool industry that focuses on comprehensive, system-level governance.
- **Pay particular attention to the intersection of AI with other vital and emerging technologies such as cybersecurity, quantum systems, blockchain, and robotics:** The convergence of these technologies can bring unique risks, but also offers strategic industry growth opportunities. As indicated above in Initiative 5, managing technology risks in isolation may lead to wasted resources and overlooked threats and opportunities.
- **Concentrate on the empirical understanding and experimentation of AI uses and technologies:** Given the complexity and automated learning aspects of AI technologies, effective governance is as much about empirical understandings of AI systems and emerging, un-designed properties as it is about by-design practices. There should be significant investment in empirical studies within the industry and independent research organisations.
- **Adopt a supply chain perspective:** Most companies employ AI-enabled technologies, components, and products within a complex supply chain, often accompanied by limited and potentially unreliable information. With the potential widespread adoption of large language models and multi-modal foundational models provided by private companies and open source, there is an increased need a to address the problematic quality of supply chain information, inscrutable models even with full transparency on artifacts, and limited control over suppliers.

Q4: Do you have suggestions on coordination of AI governance across government? Please outline the goals that any coordination mechanisms could achieve and how they could influence the development and uptake of AI in Australia.

AI governance often concerns context or sector-specific uses that involve impacted communities. In addition to issuing high-level, overarching guidelines, it is important to develop use-specific guidelines and playbooks (Initiative 1) in the context of government digital services or internal use. The coordination could also involve methods for metrics, measurement, and TEVV (Initiative 2). Additionally, there could be incentives for inventing new best practices and sharing these amongst government agencies and with industry (Initiative 3).

It is also important to adopt an experimental approach and mindset, for example through the use of a sandbox (Initiative 4) and use technology programs for strategic policy and regulation inputs (Initiative 6). Implementing an integrated responsible AI training program (Initiative 5) could help prevent the creation of risk silos. There are also benefits in coordinating efforts to assist minority groups across government agencies and use cases via minority-group-led approaches to AI (Initiative 7).

As such, it is important that a cross-government initiative is supported by a governance body, with representation from end-user groups, researchers, industry, and other relevant professionals including ethical/human rights and legal oversight.

1.3 Responses suitable for Australia

Q5: Are there any governance measures being taken or considered by other countries (including any not discussed in this paper) that are relevant, adaptable and desirable for Australia?

We note that whether countries are looking to use soft or hard law to regulate AI, these countries are involved in initiatives to shape international policy on AI and ensure interoperability of local policies. For example, the European Union has conducted important outreach activities (such as dialogue and joint initiatives with like-minded partners) through its intouchai.eu initiative, an international outreach initiative for human centric AI[4], as there is an increasing need to co-ordinate and ensure interoperability between the various policies approach to avoid the further fragmentation of AI policy initiatives and the multiplication of cyber, data and AI regulations that can make the development of AI technology challenging for organisations of any size.

Amongst the AI policy initiative, a common approach that seems to be taken is the reliance on the certification of AI systems. By embedding certification in a voluntary process (the United Kingdom and United States approach), or in a mandatory manner (the European Union approach), certification is used as a policy tool to ensure organisations adhere and follow recognised best practice for AI systems.

Indigenous populations internationally have begun governance mechanisms to ensure a space is kept privileging Indigenous worldviews as it pertains to AI[5][6]. In addition, focussing locally on Australia, the Government has a responsibility to ensure a governance structure is inclusive of Aboriginal and Torres Strait Islander voices as it pertains to AI-influenced disciplines/areas, both for cultural responsiveness and scientific rigour. Similar mechanisms are relevant, adaptable and desirable for Australia.

[4] https://digital-strategy.ec.europa.eu/en/policies/international-outreach-ai

[5] https://www.sshrc-crsh.gc.ca/funding-financement/nfrf-fnfr/stories-histoires/2023/inclusive_artificial_intelligence-intelligence_artificielle_inclusive-eng.aspx

[6] https://www.indigenous-ai.net

1.4 Target areas

Q6: Should different approaches apply to public and private sector use of AI technologies? If so, how should the approaches differ?

Nil response.

Q7: How can the Australian Government further support responsible AI practices in its own agencies?

Similar to the challenges faced by SMEs and non-AI industry sectors, it is often difficult for individual agencies to possess the requisite expertise to practice responsible AI and formulate their own use-specific guidelines. Therefore, coordination mechanisms and a more centralised effort that leverages concentrated expertise in trusted expert organisations or units are crucial. The coordination mechanisms suggested in Q4 could be useful to alleviate the expertise challenges faced by individual agencies.

Moreover, promoting a ‘government-as-exemplar’ initiative would not only coordinate the government’s implementation of responsible AI, but could also pioneer best practices that could be utilised by the broader Australian industry. This includes helping seed a responsible AI tool industry in Australia via partnerships with Australian responsible AI tool start-ups, SMEs and commercialisation efforts.

We note the current efforts led by the Department of Prime Minister and Cabinet to improve the coordination and efficient use of government expertise across the Australian Public Sector. CSIRO looks forward to assisting with this initiative.

Q8: In what circumstances are generic solutions to the risks of AI most valuable? And in what circumstances are technology-specific solutions better? Please provide some examples.

Nil response.

Q9: Given the importance of transparency across the AI lifecycle, please share your thoughts on:

a. **where and when transparency will be most critical and valuable to mitigate potential AI risks and to improve public trust and confidence in AI?**

The aspect of transparency (Initiative 7) may vary significantly, depending on whether we are discussing actual training data, model weights, code, or trusted metadata similar to the Software Bills of Materials (SBOM), which lists the key elements of the provided software. Transparency can also pertain to metrics, measurements, and TEVV methods used in product development and post-deployment monitoring.

Considering the complexity of AI technologies, stakeholders may struggle to assess the risks, even with full access to the data, model, and code. Concurrently, transparency regarding metrics, measurements, TEVV methods, and best practices is also crucial.

In addition, ensuring the integrity of data disclosed for transparency is vital. AI system development is a complex process, often integrating third-party AI tools, components, and data. Sharing data without assuring the integrity of these data may undermine trust in the shared data.

There are a number of approaches that could improve the outcome of transparency while managing sensitive information. For example, CSIRO has been tackling these challenges from the following aspects:

- In terms of metadata sharing, we are extending SBOM to AIBOM (Xia 2023b) and DataBOM to capture key elements of AI systems and are developing advanced systems to detect AI components more reliably within a system.
- To ensure the integrity of the shared data, we are using software engineering approaches to automatically detect AI components, data/model/mode provenance and dependency, cryptographically assured via provider identities and technologies like blockchain (Xu 2022).
- Regarding the sharing of metrics, measurements, TEVV and practices, our approach is to develop reusable patterns and common question banks to ensure a level of consistency during sharing.

Lastly, it is important to highlight the role of trusted third parties, particularly research organisations, which have both the expertise and trust from all sides to play a critical role in facilitating transparency requirements. This includes special access for third-party research organizations due to commercial and national security sensitivity, the integrity of the shared data/metadata, and non-regulatory data sharing.

b. mandating transparency requirements across the private and public sectors, including how these requirements could be implemented.

We refer to CSIRO's response to Positioning Australia as a leader in digital economy regulation - Automated decision making and AI regulation issues paper (see Appendix 1), where we explained how accountability and liability is used as a mechanism to ensure organisations have the appropriate levels of transparency and explainability in place:

"AI can be used to turn data about individuals against them, without individuals knowing about it or without knowing how the data was used against them (see our response to Q9 for an example in the energy sector). This is partly due to challenges in the explainability of AI decisions, the speed and scale of ADM/AI decisions and trade-secret claims from AI solution providers (Katyal and Graves 2021). This makes it very difficult for individuals to challenge decisions, insights, or inferences made about them. For that reason, some jurisdictions have been reviewing and considering appropriate liability regimes for AI systems (Tatjana 2020)."

For example, the EU has adopted the AI Liability directive which creates a rebuttable 'presumption of causality', to ease the burden of proof for victims to establish damage caused by an AI system and is also updating its product liability directive for AI[7].

We also note that in its final report (AHRC 2021) the Australian Human Rights Commission already recommended that "the Australian Government should introduce legislation that provides a rebuttable presumption to ease the burden of proof for individuals."

[7] https://www.lexology.com/library/detail.aspx?g=c153bddf-249e-40b2-89a0-24fcc94b6520

In line with the above initiatives, ISO/IEC 42001 – AI Management System, is a standard that enables the certification of AI systems and require as a key control that organisations keep appropriate records in relation to the development and use of their AI models.

Finally, we note that transparency requirements aimed directly at citizens, such as for example privacy notices, have been experiencing challenges in appropriately informing and/or protecting citizens against the misuse of their data and fostering trust[8].

Q10: Do you have suggestions for:

a. Whether any high-risk AI applications or technologies should be banned completely?

We note that:

- The EU AI Act has provided a list of prohibited high risk AI applications, and that there is a process in place that allows for this list to be updated[9].
- Regulators have cautioned against the use of AI emotion recognition[10], and Microsoft stopped selling the technology[11]. Please note that such technology, if mature and reliable, may be helpful in managing and treating mental health related issues.
- In CSIRO's discussion paper entitled Artificial Intelligence: Australia's Ethics Framework[12] (see Appendix 2), we provided from page 63-65 a detailed example of a risk assessment framework for AI, which identifies scenarios when the level of risk is unacceptable.
- Finally, we also refer to our response to Q3, Initiative 2 on developing trustworthiness metrics, measurement and TEVV methods, which can help identify high risk applications that should be banned.

b. Criteria or requirements to identify AI applications or technologies that should be banned, and in which contexts?

Nil.

[8] https://theconversation.com/94-of-australians-do-not-read-all-privacy-policies-that-apply-to-them-and-thats-rational-behaviour-96353

[9] https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

[10] https://ico.org.uk/about-the-ico/media-centre/news-and-blogs/2022/10/immature-biometric-technologies-could-be-discriminating-against-people-says-ico-in-warning-to-organisations/

[11] https://www.reuters.com/technology/microsoft-stops-selling-emotion-reading-tech-limits-face-recognition-2022-06-21/

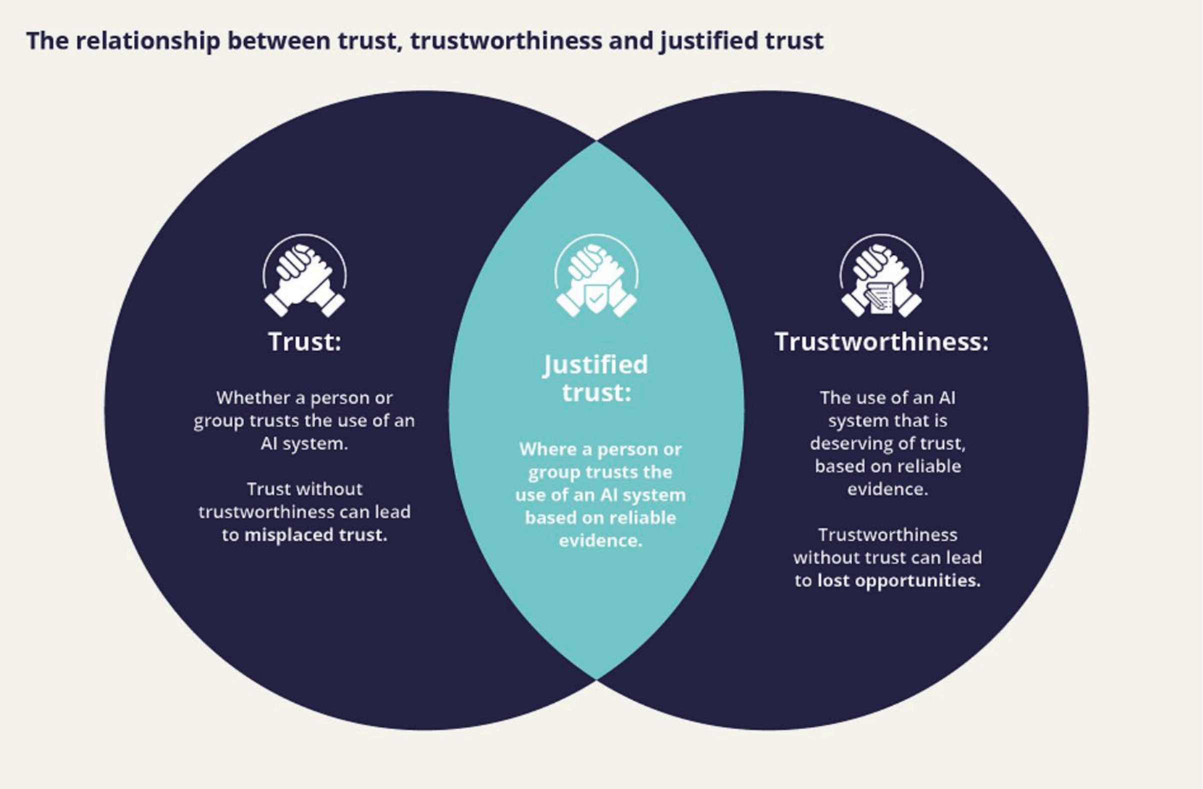
[12] https://www.csiro.au/en/research/technology-space/ai/ai-ethics-framework/

Q11: What initiatives or government action can increase public trust in AI deployment to encourage more people to use AI?

CSIRO encourages the use of internationally recognised best practice that:

1. Helps organisations implement AI responsibly, and;
2. is interoperable with other responsible AI policy initiatives so that organisations of all size can scale internationally across different jurisdictions but also easily import AI components to deliver product and services in Australia.

In line with Q3, Initiative 1, promote the creation of sector or vertical use case industry best practices, playbooks and guidelines, and Initiative 2, promote the development of trustworthiness metrics, measurement and TEVV methods, which can in turn help organisations acquire justified trust from people (see diagram below explaining UK CDEI's approach to justified trust[13]:


The relationship between trust, trustworthiness and justified trust
Trust:
Whether a person or group trusts the use of an AI system.
Trust without trustworthiness can lead to misplaced trust.
Justified trust:
Where a person or group trusts the use of an AI system based on reliable evidence.
Trustworthiness:
The use of an AI system that is deserving of trust, based on reliable evidence.
Trustworthiness without trust can lead to lost opportunities.

Create AI regulations with external, independent oversight (e.g., National Universities)*[14]*.

[13] https://www.gov.uk/government/publications/the-roadmap-to-an-effective-ai-assurance-ecosystem/the-roadmap-to-an-effective-ai-assurance-ecosystem-extended-version

[14] https://assets.kpmg.com/content/dam/kpmg/au/pdf/2023/trust-in-ai-global-insights-2023.pdf pages 31, 36.

1.5 Implications and infrastructure

Q12: How would banning high-risk activities (like social scoring or facial recognition technology in certain circumstances) impact Australia's tech sector and our trade and exports with other countries?

Nil response.

Q13: What changes (if any) to Australian conformity infrastructure might be required to support assurance processes to mitigate against potential AI risks?

Nil response.

1.6 Risk-based approaches

Q14: Do you support a risk-based approach for addressing potential AI risks? If not, is there a better approach?

When it comes to AI governance, a risk-based approach is not the only available approach. Other approaches could complement a risk-based approach, addressing specific aspects of issues to be governed.

Rights-based Approach: This method is grounded in the protection of rights such as human rights. Its key element is the establishment of an absolute threshold, which should not be compromised in an overall utilitarian manner. A risk-based approach might overlook this if the aggregated risk and benefits analysis does not incorporate such a strict threshold. For instance, CSIRO's privacy research has developed technologies to set such a hard threshold through a quantitative method, allowing trade-offs above the threshold while enforcing privacy guarantees below it[15].

Quality-based Approach: This strategy focuses on defining, verifying, validating, and maintaining quality goals and metrics for both the AI system and its development process. It involves setting quantitative benchmarks and implementing measures to ensure these benchmarks are met. Safety assurance can be considered quality-based if they set specific standards that products or services must meet. However, due to the complexity of AI technologies and their broad use cases, defining quality benchmarks and metrics for some responsible AI aspects can be challenging. For certain quality attributes such as reliability, safety, security, and privacy, an available quality-based approach can offer more assurance and trustworthiness in addition to pure risk assessment. At CSIRO, we have been developing methods for AI system-level governance, mixing quality-based methods via automated TEVV procedures with a more risk-based impact-driven approach for broader social, human value, and fairness impact assessments.

[15] https://research.csiro.au/isp/research/privacy/r4/

Principle-based Approach: This method involves creating governance measures based on a set of fundamental principles. These principles can guide the development and use of AI, providing a flexible framework that can adapt to technological advancements. Many countries, including Australia, have taken a principle-based approach, publishing high-level responsible and ethical AI frameworks that further drive the development of other approaches when operationalising these high-level principles. CSIRO, for example, developed Australia's principle-based AI ethics framework[16], using these principles to guide the different approaches during the operationalisation of AI governance (see Appendix 3 for an example of how principle-based approaches can be operationalised in different specific contexts).

Outcome-based Approach: This approach concentrates on the outcomes or results of using the AI. Instead of prescribing specific methods, processes, and quality standards, it defines the desired outcomes and associated incentives and disincentives, allowing flexibility in how these outcomes are achieved. This is particularly useful when methods, processes, and standards are rapidly evolving and challenging to set. Some product liability approaches in vertical domains are examples of an outcome-based approach. However, a purely outcome-based approach might cause significant harms before any intervention. Furthermore, due to the complex supply chain of AI systems/services and the partial responsibility of AI users, attributing appropriate responsibility to the various actors in the system can be challenging. For example, CSIRO's research has focused on tracking the provenance of data, model, and code with integrity in a supply chain setting, and has aimed to reduce the burden of proof of accountability and responsibility across the supply chain, including end users (Xu 2022).

We also note that:

a. most international policy initiatives have relied on a risk- based approach (For example: The European Union AI Act, NIST, OECD etc); and

b. existing business practices usually rely on a risk-based approach to conduct business.

Q15: What do you see as the main benefits or limitations of a risk-based approach? How can any limitations be overcome?

The primary benefits of a risk-based approach include the following:

1. A risk-based approach is best suited for scenarios where quantifiable metrics for many characteristics have not yet been established. Furthermore, by definition, a risk-based approach is especially adept at dealing with unexpected events.
2. It allows a more holistic assessment of risks and benefits, enabling trade-offs to be made within acceptable risk parameters.
3. It takes a stakeholder-driven approach that goes beyond limiting governance to impacts on end users, focusing also on the wider social, environmental, and economic impact on the broader community.

[16] https://www.industry.gov.au/publications/australias-artificial-intelligence-ethics-framework/australias-ai-ethics-principles

The main limitations of a risk-based approach include:

1. It can be challenging to quantify risks accurately and reliably, resulting in significant uncertainty in the risk assessment outcomes. Many risk assessment methods are subjective with incomplete and low-quality data.
2. If not used carefully, it could follow a purely utilitarian approach without hard boundaries, potentially neglecting edge cases and inadvertently harming minority groups.
3. To overcome the limitations of the risk-based approach, it can be combined with other approaches mentioned in Q15 for particular technological approaches, sectors, and use cases.

Q16: Is a risk-based approach better suited to some sectors, AI applications or organisations than others based on organisation size, AI maturity and resources?

As discussed in responses to Q14-Q15, risk-based approaches have their advantages and disadvantages. These can be supplemented by other strategies, each posing significant challenges to resource-constrained SMEs or organisations lacking specific expertise. The initiatives detailed in the response to Q3, particularly those related to SME profiles, use-specific guidelines, and automated TEVV methods, can help alleviate these resource and expertise challenges.

Q17: What elements should be in a risk-based approach for addressing potential AI risks? Do you support the elements presented in Attachment C?

As indicated in our response to Q9, CSIRO's discussion paper titled Artificial Intelligence: Australia's Ethics Framework (see pages 63-65) outlines a detailed example of a risk assessment framework for AI that can be leveraged and updated (as this work was conducted already in 2019) to better support the risk approach presented in Attachment C of the Safe and Responsible Artificial Intelligence (AI) in Australia discussion paper.

In terms of the elements presented in Attachment C we would draw attention to current international standardisation work on AI and related certification initiatives, including specifically:

- ISO 42005 on AI impact assessment that provide the details and guidance on how to conduct an AI impact assessment that can be incorporated in and is aligned with existing business practices.
- ISO 42001 that covers on the specific risk requirements presented in Attachment C and provide further details and guidance to implement them. For example, on documentation, ISO 42001 imposes a requirement that documentation is in place and kept and explain what resources documentation is required (e.g., data, tooling, human resource etc.) it also imposes requirements for each stage of the AI system lifecycle including monitoring.

There are other ISO standards that also specifically provide guidance on controllability, explainability and transparency methods, process and requirements for AI systems.

Additionally, in reference to Q3, which discusses the sources of AI risk such as technological approaches, use cases, and AI capabilities, it would be beneficial to incorporate these as distinct dimensions within a risk-based approach.

Attachment C of the Safe and Responsible Artificial Intelligence (AI) in Australia discussion paper presents a good start. We note and refer to current challenges in privacy laws when it comes to relying on using user notification as a transparency mechanism.

Q18: How can an AI risk-based approach be incorporated into existing assessment frameworks (like privacy) or risk management processes to streamline and reduce potential duplication?

CSIRO's experts together with other Australian experts have developed and provided guidance specifically on this point as part of ISO/IEC 42005 on AI Impact Assessment[17]. This guidance specifically explain how an AI Impact assessment can be embedded in existing practices and aligned with privacy impact assessment, human rights impact assessments and such other assessments to reduce duplication and facilitate the re-assessment of AI systems. To date, this guidance has been accepted and has been very well received by other international standards experts on AI.

The answers to Q14 and Q15 also provided suggestions on how risk-based approaches can be integrated with other frameworks which may have used an alternative approach or mixed approach.

Q19: How might a risk-based approach apply to general purpose AI systems, such as large language models (LLMs) or multimodal foundation models (MFMs)?

Similar to other AI technologies, LLMs and MFMs serve as technological components for creating a wide range of products and services. However, there are several qualitative and quantitative differences that these models present concerning responsible AI and a risk-based approach:

1. Unlike smaller AI models, training LLMs/MFMs often requires significant computational resources, data access, and expertise. As a result, most organisations only utilise third-party or open source LLMs/MFMs as key components of their products and services, resulting in less information and control over these models' risks. Many traditional AI governance approaches designed for self-training AI models may not apply.
2. LLMs/MFMs are highly capable, generic models that cannot be easily pre-tuned to mitigate risks for particular use cases. Organisations using LLMs/MFMs to build products and services will need to take some generic but unsafe models and mitigate the risks at the system level, as opposed to the model level. However, system-level practices are still evolving, as many technical AI governance practices focus on the model level, such as data governance and model training governance.
3. LLMs/MFMs are substantially more difficult to understand, and control compared to smaller AI models. Moreover, understanding the sources of training data is challenging due to the data's size and the commercial sensitivity often associated with this information from model providers.
4. LLMs/MFMs are easily accessible via numerous consumer tools such as separate chatbots, image generators, and common tools like Microsoft Office and Photoshop. The ease of access, including through shadow IT, makes governing the responsible use of LLM/MFM-based products very challenging.

[17] https://www.iso.org/standard/44545.html

Given these differences, specific governance approaches (risk-based and other complementary approaches) for LLMs/MFMs need to be developed building on the governance approach used for AI systems. For instance, CSIRO has initiated work to tailor its design pattern-based responsible AI approaches to foundation models, proposing a reference architecture and responsible AI considerations (Lu 2023b, Lu 2023c).

Similar to AI systems, the harms that LLMs/MFMs can cause depend on their use case. Organisations often use risk-based approaches to manage risks across their businesses, such as privacy assessments, human rights assessments, and procurement assessments, and these risk-based assessments can incorporate LLM/MFM-specific characteristics and apply them to address use case risks.

Q20: Should a risk-based approach for responsible AI be a voluntary or self-regulation tool or be mandated through regulation? And should it apply to:

a. **public or private organisations or both?**

b. **developers or deployers or both?**

Nil Response.

References

AHRC. 2021. “Human Rights and Technology: Final Report.” Australian Human Rights Commission (AHRC).

Boming Xia, Qinghua Lu, Harsha Perera, Liming Zhu, Zhenchang Xing, Yue Liu, Jon Whittle. 2023a. “Towards Concrete and Connected AI Risk Assessment (C2AIRA): A Systematic Mapping Study.” 2023 ACM/IEEE 2nd International Conference on AI Engineering (CAIN’2023).

Boming Xia, Tingting Bi, Zhenchang Xing, Qinghua Lu, Liming Zhu. 2023b. “An Empirical Study on Software Bill of Materials: Where We Stand and the Road Ahead.” 2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE’2023).

CSIRO. 2022. Responsible AI Pattern Catalogue. CSIRO’s Data61. https://research.csiro.au/ss/science/projects/responsible-ai-pattern-catalogue/.

Evas Tatjana. 2020. “Civil Liability Regime for Artificial Intelligence: European Added Value Assessment.” LU: European Parliamentary Research Service. https://data.europa.eu/doi/10.2861/737677.

Jonas Schuett. 2023. Defining the scope of AI regulations. Law, Innovation and Technology 1–23. https://doi.org/10.1080/17579961.2023.2184135.

Mills, M.C., Rahal, C. The GWAS Diversity Monitor tracks diversity by disease in real time. *Nat Genet* **52**, 242–243 (2020). https://doi.org/10.1038/s41588-020-0580-y

Qinghua Lu, Yuxiu Luo, Liming Zhu, Mingjian Tang, Xiwei Xu, Jon Whittle. 2023a. “Developing Responsible Chatbots for Financial Services: A Pattern-Oriented Responsible AI Engineering Approach.” IEEE Intelligent Systems.

Qinghua Lu, Liming Zhu, Xiwei Xu, Zhenchang Xing, Jon Whittle. 2023b. “Towards Responsible AI in the Era of ChatGPT: A Reference Architecture for Designing Foundation Model-based AI Systems.” https://arxiv.org/abs/2304.11090.

Qinghua Lu, Liming Zhu, Xiwei Xu, Zhenchang Xing, Jon Whittle. 2023c. “A Taxonomy of Foundation Model based Systems for responsible-AI-by-design.” https://arxiv.org/abs/2305.05352.

Shum, B. et al. 2022. The inequity of targeted cystic fibrosis reproductive carrier screening tests in Australia. Prenat Diagn. 2023 Jan;43(1):109-116. doi: 10.1002/pd.6285. Epub 2022 Dec 15. PMID: 36484552.

Silvera, D. et al (2022) Framework for Artificial Intelligence enabled Assistive Technology as

Supports under the National Disability Insurance Scheme: Final Report. https://ndis.gov.au/about-us/research-and-evaluation/market-stewardship-and-employment/markets-and-innovations-research

Sonia Katyal and Charles Graves. 2021. “From Trade Secrecy to Seclusion.” *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3760123.

Sung Une Lee, Harsha Perera, Boming Xia, Yue Liu, Qinghua Lu, Liming Zhu, Olivier Salvado, Jon Whittle. 2023. “QB4AIRA: A Question Bank for AI Risk Assessment.” https://arxiv.org/abs/2305.09300.

Walter, M. (2018) ‘The Voice of Indigenous Data: Beyond the Markers of Disadvantage’ First Things First, Griffith Review, (60).

Xiwei Xu, Chen Wang, Zhen Wang, Qinghua Lu, and Liming Zhu. 2022. Dependency tracking for risk mitigation in machine learning (ML) systems. 44th International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP '22). https://doi.org/10.1145/3510457.3513058.

Appendix 1: Positioning Australia as a Leader in Digital Economy Regulation – Automated Decision Making and AI Regulation Issues Paper

Australia's National
Science Agency

Positioning Australia as a leader in digital economy regulation - Automated decision making and AI regulation issues paper

CSIRO Submission 22/792

May 2022

Main Submission Author(s):
Liming Zhu, Research Director, CSIRO Data61

Enquiries should be addressed to:
Ms Liz Yuncken
CSIRO Government Relations
GPO Box 1700 Canberra 2601
T 02 6218 3547
E mplo@csiro.au

Contents

Introduction

CSIRO welcomes the opportunity to provide comment on the Digital Technology Taskforce Issues Paper - Positioning Australia as a leader in digital economy regulation - Automated Decision Making and AI Regulation.

This submission addresses questions in the discussion paper that relate to CSIRO's scientific and technological expertise.

As Australia's national science agency, CSIRO is at the centre of solving Australia's greatest challenges through innovative science and technology. CSIRO contributes to inclusive, ethical, safe, and secure AI adoption in Australia through research in areas such as advanced AI technologies, AI engineering, responsible AI and innovation, human-AI collaborative intelligence, regulation technology, privacy, and cybersecurity of AI systems. CSIRO hosts the National AI Centre as part of Australia's AI Action Plan and helps Australian industry adopt AI in strategically important sectors such as manufacturing, energy, agriculture, critical minerals, digital services, and defence.

CSIRO welcomes the opportunity to discuss these matters in more depth with the taskforce. Please refer to the contact details on the cover page.

CSIRO response

1. What are the most significant regulatory barriers to achieving the potential offered by AI and ADM? How can those barriers be overcome?

Based on our research and interactions with industry, CSIRO considers the following regulatory barriers to be a hinderance to achieving the potential offered by artificial intelligence (AI) and automated decision-making (ADM):

1. **Regulatory uncertainty and regulatory fragmentation:** are often listed amongst some of the most significant barriers to AI adoption and expansion by industry. In the response at question 10, we outline how AI standardisation initiatives can help enable interoperability between the different regulatory approaches and how standards have been leveraged to achieve such interoperability in the past. To enable further regularity clarity and certainty, it is also worth noting that independent and reputable science institutions can, and are, well placed to develop concrete (social-) technical measurements of the more general requirements set out in the standards (NIST 2022). For AI this is particularly relevant as precise measurements can assist stakeholders to manage conflicts between requirements, such as accuracy, fairness, and privacy.

2. **Liability:** AI can be used to turn data about individuals against them, without individuals knowing about it or without knowing how the data was used against them (see our response to question 9 for an example in the energy sector). This is partly due to challenges in the explain-ability of AI decisions, the speed and scale of ADM/AI decisions and trade-secret claims from AI solution providers (Katyal and Graves 2021). This makes it very difficult for individuals to challenge decisions, insights, or inferences made about them. For that reason, some jurisdictions have been reviewing and considering appropriate liability regimes for AI systems (Tatjana 2020). Similarly, we note that in its final report (AHRC 2021) the Australian Human Rights Commission recommended that “the Australian Government should introduce legislation that provides a rebuttable presumption that, where a corporation or other legal person is responsible for making a decision, that legal person is legally liable for the decision regardless of how it is made, including where the decision is automated or is made using artificial intelligence”.

3. **Operationalisation**: Our research has identified that currently there is insufficient guidance on how to appropriately operationalise the relevant high-level principles and principle-based regulations into the end-to-end development and monitoring of AI systems (Zhu et al. 2022). There is an increasing demand for tools that help technologists, managers and other stakeholders check compliance and conformance with best practices, standards, policies, and regulations. For example, tools such as R4 [software] (CSIRO 2022a) can help technologists and stakeholders manage re-identification risks, and a responsible AI pattern catalogue (CSIRO 2022b) can help technologists share and reuse best practices.

4. **Data sharing:** As AI and ADM systems require quality data, it is important to facilitate data asset creation and data sharing initiatives that can be trusted and benefit Australians. However, in CSIRO's experience data asset creation and data sharing is restricted owing to a range of IP, privacy, and confidentiality risks. In our response to question 6 of this document, we explain that to address this challenge data custodians could be provided with clearer direction and guidance on the controlled release of data. For example, this is a particular need that has been identified for the Australian energy sector as energy data custodians such as the Australian Energy Market Operator (AEMO), the Australian Energy Regulator (AER) and network businesses could benefit from improved regulatory guidance on when and how to release energy data while managing privacy, confidentiality, and security risks (https://esb-post2025-market-design.aemc.gov.au/32572/1630275857-esb-data-strategy-july-2021.pdf).

   In some circumstances, additional regulations may also be required to facilitate improved access to data necessary for maintaining system security, reliability, and efficiency. CSIRO has expertise in this area and previously hosted and provides technical advice to the Data Standard Body of the Consumer Data Right, releasing data sharing, security and consent management standards for the banking and energy sectors. CSIRO is also working with key energy data custodians such as the AEMO and network businesses to provide best practice risk management and de-identification controls to support controlled data release programs.

5. **Human involvement:** With the increased complexity of AI systems and the issue of human over-reliance on technology (otherwise known as automation bias and automation complacency see: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6534180/), in CSIRO's experience it is becoming increasingly important to understand what human involvement is possible or desirable and to carefully design "meaningful" human involvement. For example, design that: (1) considers human competency and its limits; and (2) identifies the most suitable mechanisms (either human, technical or a combination of both) to deliver better control over dynamic systems like AI.

   We note there is often a general assumption that fully automated systems present higher risks than system with humans in the loop, as the latter provides for human accountability. Yet, this approach can, and has, resulted in incentivising organisations to superficially add a human in the loop and use them as a liability and accountability "sponge." For example, the European Union's General Data Protection Regulation (EU GDPR) imposes additional requirements on AI systems that are fully automated in comparison to AI systems where there are humans in the loop. Similarly, the US Food and Drug Administration is requiring approval of systems that are automating decisions but not of systems that are only supporting decisions made by health care professionals (as they are not characterised as medical devices). To refine this approach research is being conducted: (1) in the field of Human Centred Design and Human Computer interaction to help better understand the human behaviours, and capabilities constraints when using AI; and (2) on technical tools that can better enable human oversights of AI. The International Organization for

Standardization (ISO) is currently developing work on controllability of automated AI systems, see https://www.iso.org/standard/83012.html?browse=tc).

2. Are there specific examples of regulatory overlap or duplication that create a barrier to the adoption of AI or ADM? If so, how could that overlap or duplication be addressed?

Many regulations relevant to AI are at the State Government level and this can make the development of uniform and nationally scalable approaches challenging. This is the case for example for AI and ADM for the energy sector and for autonomous vehicles.

3. What specific regulatory changes could the Commonwealth implement to promote increased adoption of AI and ADM? What are the costs and benefits (in general terms) of any suggested policy change?

As noted in the response to question 1, if AI systems are used to automate decisions at scale that can impact individuals, then CSIRO suggests that co-design should be an essential feature of the AI development process.

Additionally, research (Smuha 2021) has shown that it is important to have “(1) public oversight mechanisms to increase accountability, including mandatory impact assessments with the opportunity to provide societal feedback; (2) public monitoring mechanisms to ensure independent information gathering and dissemination about AI’s societal impact; and (3) the introduction of procedural rights with a societal dimension, including a right to access to information, access to justice, and participation in public decision-making on AI, regardless of the demonstration of individual harm”.

For example, in the energy sector, the Energy Security Board (ESB) data strategy and the Consumer Data Right for Energy outline critical regulatory changes toward facilitating improved energy data access while protecting consumer privacy and confidentiality. The costs of these changes will primarily be in developing improved capacity within data custodian organisations for controlled data release. The benefits are substantial both in more efficient system operations and avoided capital expenditure.

4. Are there specific examples where regulations have limited opportunities to innovate through the adoption of AI or ADM?

Research has identified that existing regulations are impacting AI/ADM indirectly by requiring, prohibiting, encouraging, and discouraging human involvement in decision making (Crootof, Kaminski, and Prince 2022). In CSIRO’s experience, simply having human involvement in the loop for AI/ADM solutions does not solve the problem because, without “appropriate” human involvement, the hybrid system may introduce more risks, provide perverted incentives to include a human in the loop to avoid certain regulation, add inefficiency, or delay AI/ADM adoption. Sometimes meaningful human involvement may only be possible at design time and at a high level rather than directly at run time. The appropriate time and level of human involvement in autonomous operations in the marine context is one such example.

In the energy sector, CSIRO's experience is that the Power of Choice reforms and the contestability of metering services has introduced new challenges to network businesses and their sub-processors accessing the smart meter data necessary for AI and ADM in network operations. While the reforms addressed the potential for innovation and competition in retail services, they may not have sufficiently considered the important role of smart meter data in network AI and ADM. The ESB Data Strategy and associated reforms are attempting to address this limitation.

5. Are there opportunities to make regulation more technology neutral, so that it will more apply more appropriately to AI, ADM and future changes to technology?

Standardisation is used in the European Union as a powerful tool to support and enable both the ongoing reliance and development of technology neutral laws. The Executive Vice-President for a Europe Fit for the Digital Age, Margrethe Vestager highlights: "Ensuring that data is protected in artificial intelligence or ensuring that mobile devices are secure from hacking, rely on standards and must be in line with EU democratic values. In the same way, we need standards for the roll-out of important investment projects, like hydrogen or batteries, and to valorise innovation investment …"(https://ec.europa.eu/growth/news/new-approach-enable-global-leadership-eu-standards-promoting-values-and-resilient-green-and-digital-2022-02-02_en).

Technical standardisation norms and processes have become prominent features of post-national policy making in Europe, since the development of the 'new approach to technical harmonisation' in 1985 (the New Approach. One of the main changes was a move away from laws detailing all the technical and administrative requirements, to a New Approach, which restricts the content of legislation to 'essential requirements', leaving the technical details to European harmonised standards (see https://sesei.eu/european-standardization/new-approach-legislation/).

6. Are there actions that regulators could be taking to facilitate the adoption of AI and ADM?

Promoting collaboration between research organisations and regulators and/or a dedicated AI regulator (such as the AI Safety Commissioner proposed by the Australian Human Rights Commission in its Human Rights and Technology Final Report) can be a powerful tool to help develop, and promote, science-informed and data-driven regulatory guidance that leverages state-of-the-art scientific research. Examples from the United Kingdom are the collaborations developed by the Alan Turing Institute with the Information Privacy Commissioner on AI explain-ability (https://www.turing.ac.uk/blog/project-explain-enters-its-next-phase), and with the Financial Conduct Authority on AI transparency (https://www.turing.ac.uk/news/ai-transparency-financial-services).

On the basis that AI systems require quality data to be successful, there is a need to facilitate data sharing initiatives that can be trusted and benefit Australians. For that reason, one of the main actions regulators could take to facilitate AI and ADM, is to provide custodians of government-held data with clear direction and guidance on the controlled release of data., Regulators could also assist by specifying data standards and interchange protocols. This is a particular need that has been identified for the Australian energy sector and would provide a strong foundation for scalable AI and ADM innovation in the energy sector.

7. Is there a need for new regulation or guidance to minimise existing and emerging risks of adopting AI and ADM?

CSIRO is working on projects (https://research.csiro.au/ri/an-operationalised-guideline-for-responsible-ai/) that aim to operationalise the eight high-level Responsible AI principles identified by *Australia's Artificial Intelligence Ethics Framework* (Dawson et al. 2019). CSIRO is taking an end-to-end AI life cycle approach to go beyond the typical AI/ADM algorithm-driven techniques to include stakeholder consultation, data collection, requirements, system design, monitoring, process assurance and organizational maturity (Lu et al. 2022)

CSIRO is also addressing the need to develop concrete (social-) technical measurements of the more general requirements set out in frameworks, principles, and standards (NIST 2022) and their integration into existing organisational processes for AI system development, risk management, ethics approval and other compliance processes.

Through this work, CSIRO has identified there is a need to develop reference architectures, reference implementations, and example implementations for common/critical scenarios and emerging technologies. This will particularly help SMEs to comply with current and future policies, standards, and regulations.

We also note that while AI and ADM are highly susceptible to foreign interference, we believe that existing regulation and guidance is sufficient to support agencies to increase their resilience against it. For example, CSIRO's efforts against foreign interference are aligned with the Universities Foreign Interference Taskforce guidelines, and work effectively to protect our people and our research.

8. Would increased automation of decision making have adverse implications for vulnerable groups? How could any adverse implications be ameliorated?

Research highlights that under-representation of minority and marginalised groups in the datasets used to train AI/ADM models, lack of diversity in the systems development teams, and non-adherence to principles and practices of diversity and inclusiveness in the overall deployment, adoption, and governance of these systems, can cause digital redlining, discrimination, and algorithmic oppression (https://www.chiefscientist.gov.au/news-and-media/why-we-need-think-about-diversity-and-ethics-ai).

An increase in ADM/AI does not necessarily imply that impacts on the vulnerable groups will increase, however they will be systematised, meaning that unfair and unjust decisions made (either by ADM/AI or by humans based on the results obtained from ADM/AI) may have “built in” adverse implications on these groups. In CSIRO's view, adverse implications for vulnerable groups are likely to occur when:

1. The responsibility of deciding what benefits society is missing in the practical application of not only the ethical concerns but also legal, economic and cultural concerns. This is because when AI/ADM automate for the majority, they can leave the minorities behind.
2. There is a lack of diversity in the datasets used for training ADM/AI algorithms. This in turn leads to data blindness and creates a capability challenge for AI and ADM to make/provide accurate or appropriate decisions/insights for minorities including vulnerable individuals.
3. There is a lack of diversity and inclusion in both the development and the adoption processes of AI/ADM systems, which can in turn lead to potentially unsafe or inappropriate use of AI/ADM systems.

Recently there have been many initiatives to propose frameworks for AI governance for both the development and deployment of AI systems. Most of these initiatives focus on areas of transparency, accountability, bias, privacy, non-discrimination, and other generally agreed upon values from over 100 sets of AI Ethics principles (Fjeld et al. 2020), with most having some focus on AI governance. CSIRO suggests that the development of concrete guidance on AI governance, measurements, and implementation, that maps to international standards and best practices while incorporating diversity and inclusion requirements should be considered. AI-enabled systems research and development must also be informed and shaped by diversity and inclusion principles in order to help us build trust and confidence in these systems.

9. Are there specific circumstances in which AI or ADM are not appropriate?

CSIRO's Discussion Paper *Artificial Intelligence, Australia's Ethics Framework* (Dawson et al. 2019), highlights an example of a Risk Assessment Framework for AI Systems, and examines: (1) the probability of risk together with their consequences; (2) the factors that can cause on AI application to contain more risk; and (3) the variety of actions that can be taken to mitigate risks, which in turn enabled us to identify areas of extreme level risk where organisations would usually consider the risks as unacceptable under existing laws.

We also note that in 2021 the Australian Human Rights Commission (AHRC 2021) recommended law reform to provide better human rights and privacy protection regarding the development and use of facial recognition and other biometrics technologies (Recommendations 19, 21), and a moratorium on the use of biometric technologies in high-risk decision making until such protections are in place (Recommendation 20). A similar approach has been taken by American and European legislators.

More generally, where the use of AI and ADM can cause harm, and the output of the AI system is not transparent nor explainable, its deployment is not appropriate. For example, when harm occurs and AI or ADM are used, organisations have been required to provide evidence as to how the automated action/decision was taken or made in compliance with the laws. For that reason, AI solutions that can cause harm to individuals need to be explainable and transparent to demonstrate and ensure accountability with existing laws. To enable such applications to operate without being transparent or explainable would only increase customer mistrust and also the illegal use of such technologies.

For example, with AI, electricity retailers will have an unprecedented ability to micro-target their service offerings to households and businesses. Care should be taken to ensure that such micro-targeting is done in a transparent and explainable manner to enable individuals to challenge insights or decisions made about them, and effectively ensure that they are not limited/imprisoned by their data and are not prevented from being given fair access to essential services such as electricity. Similarly, concerns have been raised around the ability of large electricity retailers to use virtual power plants and AI and ADM to manipulate market pricing for financial benefit. These perverse behaviours can be avoided through prudent regulation.

10. Are there international policy measures, legal frameworks or proposals on AI or ADM that should be considered for adoption in Australia? Is consistency or interoperability with foreign approaches desirable?

Research highlights two key reasons interoperability with foreign approaches is essential to enable innovation and the broader adoption of AI technology (Lu et al. 2022):

(1) AI systems largely depend on data. For example, an AI system may use data from multiple jurisdictions for its training or when in production. Without interoperability, and if an AI system processes data from multiple locations, different and potentially conflicting requirements and controls could apply to the data and also to the AI systems that is processing it. We note the different responses adopted by the US and Australian regulators when an AI system processes illegally obtained personal data.

(2) AI systems often only form part of the many components of supply chains. For example, a foreign AI system can be a component of the supply chain for a product or service used in Australia, or an Australian AI system can be a component of the supply chain for the delivery of a product or service overseas. In each case, the lack of interoperability between jurisdictions is likely to constitute a substantial barrier to the fast development and adoption of AI technology and to the ability of Australian companies to operate in global markets.

Industry Standards, are international policy measures that can enable interoperability.

The work of ISO/IEC JTC1/SC42 on the standardisation of AI system is one of the largest international initiatives on AI to date, with a total of 50 countries involved (34 participants including Australia and 16 observers) and has been recognised as an important tool to enable interoperability.

CSIRO's Discussion Paper, *Artificial Intelligence, Australia's Ethics Framework* (Dawson et al. 2019) outlines a toolkit for ethical AI, which includes and refers to industry standards and certification and explains the role AI Standards can play. We also refer to *Standards Australia Roadmap: Making Australia's Voice Heard* (SA 2020).

Standards can be used either on a voluntary or mandatory basis, and they can provide a bridge between the different AI regulatory approaches adopted internationally. The US Standards Bodies, NIST, is developing an AI Risk Framework that is voluntary (NIST 2022), while the European Commission is looking to the standards work to support its proposed EU AI Act, and specifically the 'CE' marking [1]of high-risk applications mandated by the Act (EU Commission 2021).

Some examples where Australia is successfully using standards to enable interoperability include:

- **Privacy:** ISO/IEC 27701 (Privacy Information Management), which contains an annex mapped to local Australian Privacy Law requirements. This Standard provides Australian businesses and the community with an interoperable privacy risk management framework that aligns with local requirements and those of EU's General Data Protection Regulation (GDPR), APEC's Cross Border Privacy Rules (CBPR) and other regional privacy frameworks.
- **Critical technology**: We also note that under actions 32 and 33 of its International Cyber and Critical Technology Engagement Strategy, the Australian Government is using critical technology standards to foster interoperability, innovation, transparency, diverse market, and security by design. The Strategy also notes the importance of international standards for emerging and critical technologies to ensure (1) economic competition and geopolitical security and (2) products, services and systems are safe, consistent, reliable, and interoperable, and (3) consistency for consumers, provide confidence in the safety and reliability of products, and when internationally harmonised, assist in reducing barriers to international trade.

Participation in international fora, such as ISO, Institute of Electrical and Electronics Engineers, and International Telecommunication Union, may help Australia to work with other countries to ensure that democratic values are reflected and promoted by international standards and other international policy instruments. Likewise, focusing on interoperability with key international allies in international fora may assist to ensure Australia is well placed to take up opportunities for collaboration and does not face barriers such as incompatible technology, or untenable security settings. For example, the US has already demonstrated its expectations of its foreign allies in other technical arenas, such as in its expectations of network hygiene in its Foreign Acquisition Regulations.

References

AHRC. 2021. "Human Rights and Technology: Final Report." Australian Human Rights Commission (AHRC).

Crootof, Rebecca, Margot Kaminski, and W. Nicholson Prince. 2022. "Humans in the Loop." *Vanderbilt Law Review*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4066781.

CSIRO. 2022a. *Re-Identification Risk Ready Reckoner (R4) [Software]*. CSIRO's Data61. https://data61.csiro.au/en/Our-Research/Our-Work/R4.

[1] CE is the abbreviation of "conformité européenne" (French for "European conformity"). The letters 'CE' appear on many products traded on the extended Single Market in the European Economic Area (EEA). They signify that products sold in the EEA have been assessed to meet high safety, health, and environmental protection requirements. Source: https://ec.europa.eu/growth/single-market/ce-marking_en

CSIRO. 2022b. *Responsible AI Pattern Catalogue*. CSIRO's Data61. https://research.csiro.au/ss/science/projects/responsible-ai-pattern-catalogue/.

Dawson, D, E Schleiger, J Horton, J McLaughlin, C Robinson, G Quezada, J Scowcroft, and S Hajkowicz. 2019. "Artificial Intelligence - Australia's Ethics Framework - A Discussion Paper." CSIRO's Data61.

Fjeld, Jessica, Nele Achten, Hannah Hilligoss, Adam Nagy, and Madhulika Srikumar. 2020. "Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI." Berkman Klein Center for Internet & Society. https://www.ssrn.com/abstract=3518482.

Katyal, Sonia, and Charles Graves. 2021. "From Trade Secrecy to Seclusion." *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3760123.

Lu, Qinghua, Liming Zhu, Xiwei Xu, Jon Whittle, and Zhenchang Xing. 2022. "Towards a Roadmap on Software Engineering for Responsible AI." In *1st International Conference on AI Engineering - Software Engineering for AI (CAIN)*. http://arxiv.org/abs/2203.08594.

NIST. 2022. "AI Risk Management Framework: Initial Draft." National Institute of Standards and Technology (NIST).

SA. 2020. "An Artificial Intelligence Standards Roadmap: Making Australia's Voice Heard." Standards Australia (SA).

Smuha, Nathalie A. 2021. "Beyond the Individual: Governing AI's Societal Harm." *Internet Policy Review* 10 (3). https://doi.org/10.14763/2021.3.1574.

Tatjana, Evas. 2020. "Civil Liability Regime for Artificial Intelligence: European Added Value Assessment." LU: European Parliamentary Research Service. https://data.europa.eu/doi/10.2861/737677.

Zhu, Liming, Xiwei Xu, Qinghua Lu, Guido Governatori, and Jon Whittle. 2022. "AI and Ethics—Operationalizing Responsible AI." In *Humanity Driven AI*, edited by Fang Chen and Jianlong Zhou, 15–33. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-72188-6_2.

As Australia’s national science agency and innovation catalyst, CSIRO is solving the greatest challenges through innovative science and technology.

CSIRO. Unlocking a better future for everyone.

www.csiro.au

Appendix 2: Artificial Intelligence: Australia’s Ethics Framework

Artificial Intelligence

Australia’s Ethics Framework

A Discussion Paper

CSIRO

DATA 61

Australian Government
Department of Industry,
Innovation and Science

Citation

Dawson D and Schleiger E*, Horton J, McLaughlin J, Robinson C∞, Quezada G, Scowcroft J, and Hajkowicz S† (2019) Artificial Intelligence: Australia's Ethics Framework. Data61 CSIRO, Australia.

*Joint first authors
∞CSIRO Land and Water
†Corresponding author

Copyright

Important disclaimer

CSIRO advises that the information contained in this publication comprises general statements based on scientific research. The reader is advised and needs to be aware that such information may be incomplete or unable to be used in any specific situation. No reliance or actions must therefore be made on that information without seeking prior expert professional, scientific and technical advice. To the extent permitted by law, CSIRO (including its employees and consultants) excludes all liability to any person for any consequences, including but not limited to all losses, damages, costs, expenses and any other compensation, arising directly or indirectly from using this publication (in part or in whole) and any information or material contained in it.

Acknowledgements

This study was funded by the Australian Government Department of Industry Innovation and Science. The authors would like to express gratitude to the experts from industry, government and community organisations who formed the steering committee that guided the research.

The authors are grateful to the many researchers from CSIRO, Data61 and universities who shared their expertise. In particular, the authors would like to thank the reviewers, including Professor Toby Walsh (Data61 and University of New South Wales), Professor Lyria Bennett Moses (University of New South Wales), Professor Julian Savulescu (University of Oxford and Monash University) and Dr Cecilia Etulain (Alfred Hospital).

The authors would also like to thank the individuals who attended the consultative workshops in Sydney, Brisbane, Melbourne and Perth and those people who attended the Data61 Live Conference in Brisbane in September 2018. By sharing your knowledge and perspective on artificial intelligence you have helped make this report possible.

Accessibility

CSIRO is committed to providing web accessible content wherever possible. If you are having difficulties with accessing this document please contact csiroenquiries@csiro.au.

Artificial Intelligence: Australia’s Ethics Framework

Public Consultation

Artificial Intelligence (AI) has the potential to increase our well-being; lift our economy; improve society by, for instance, making it more inclusive; and help the environment by using the planet's resources more sustainably. For Australia to realise these benefits however, it will be important for citizens to have trust in the AI applications developed by businesses, governments and academia. One way to achieve this is to align the design and application of AI with ethical and inclusive values.

Consultation Approach

The purpose of this public consultation is to seek your views on the discussion paper developed by Data61: ***Artificial Intelligence: Australia’s Ethics Framework.*** Your feedback will inform the Government’s approach to AI ethics in Australia.

As part of this consultation, the Department of Industry, Innovation and Science welcomes written submissions, which will close on Friday, 31 May 2019.

Please note that comments and submissions will be published on the department’s website unless, on submission, you clearly indicate that you would like your comments to be treated as confidential.

Questions for consideration:

1. Are the principles put forward in the discussion paper the right ones? Is anything missing?
2. Do the principles put forward in the discussion paper sufficiently reflect the values of the Australian public?
3. As an organisation, if you designed or implemented an AI system based on these principles, would this meet the needs of your customers and/or suppliers? What other principles might be required to meet the needs of your customers and/or suppliers?
4. Would the proposed tools enable you or your organisation to implement the core principles for ethical AI?
5. What other tools or support mechanisms would you need to be able to implement principles for ethical AI?
6. Are there already best-practice models that you know of in related fields that can serve as a template to follow in the practical application of ethical AI?
7. Are there additional ethical issues related to AI that have not been raised in the discussion paper? What are they and why are they important?

Closing date for written submissions: Friday 31 May 2019

Email: artificial.intelligence@industry.gov.au

Website: https://consult.industry.gov.au/

Mail: Artificial Intelligence
Strategic Policy Division
Department of Industry, Innovation and Science
GPO Box 2013, Canberra, ACT, 2601

Executive summary

The ethics of artificial intelligence are of growing importance. Artificial intelligence (AI) is changing societies and economies around the world. Data61 analysis reveals that over the past few years, 14 countries and international organisations have announced AU$ 86 billion for AI programs. Some of these technologies are powerful, which means they have considerable potential for both improved ethical outcomes as well as ethical risks. This report identifies key principles and measures that can be used to achieve the best possible results from AI, while keeping the well-being of Australians as the top priority.

Countries worldwide are developing solutions. Recent advances in AI-enabled technologies have prompted a wave of responses across the globe, as nations attempt to tackle emerging ethical issues (Figure 1). Germany has delved into the ethics of automated vehicles, rolling out the most comprehensive government-led ethical guidance on their development available [1]. New York has put in place an automated decisions task force, to review key systems used by government agencies for accountability and fairness [2]. The UK has a number of government advisory bodies, notably the Centre for Data Ethics and Innovation [3]. The European Union has explicitly highlighted ethical AI development as a source of competitive advantage [4].

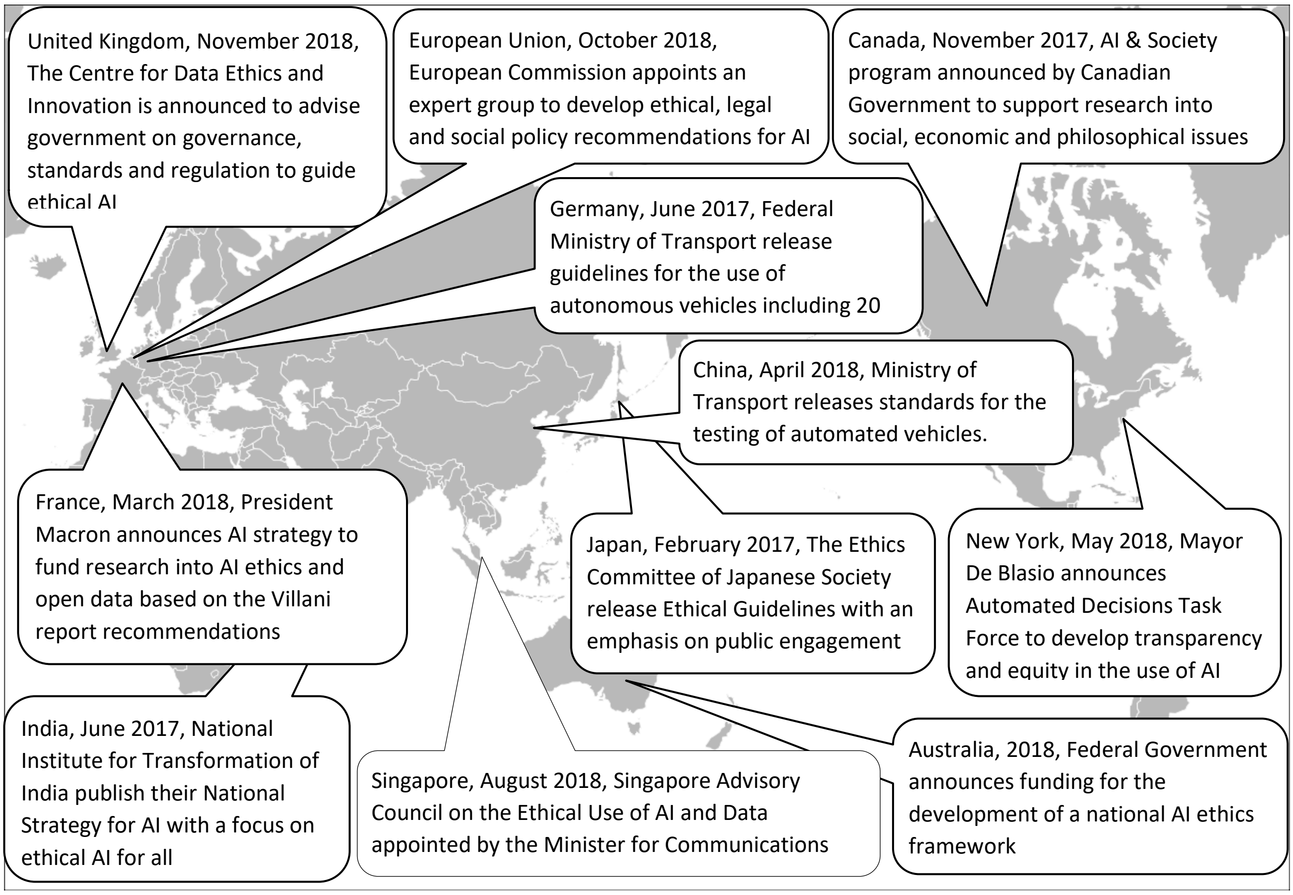

United Kingdom, November 2018, The Centre for Data Ethics and Innovation is announced to advise government on governance, standards and regulation to guide ethical AI
European Union, October 2018, European Commission appoints an expert group to develop ethical, legal and social policy recommendations for AI
Canada, November 2017, AI & Society program announced by Canadian Government to support research into social, economic and philosophical issues
Germany, June 2017, Federal Ministry of Transport release guidelines for the use of autonomous vehicles including 20
China, April 2018, Ministry of Transport releases standards for the testing of automated vehicles.
France, March 2018, President Macron announces AI strategy to fund research into AI ethics and open data based on the Villani report recommendations
Japan, February 2017, The Ethics Committee of Japanese Society release Ethical Guidelines with an emphasis on public engagement
New York, May 2018, Mayor De Blasio announces Automated Decisions Task Force to develop transparency and equity in the use of AI
India, June 2017, National Institute for Transformation of India publish their National Strategy for AI with a focus on ethical AI for all
Singapore, August 2018, Singapore Advisory Council on the Ethical Use of AI and Data appointed by the Minister for Communications
Australia, 2018, Federal Government announces funding for the development of a national AI ethics framework

Figure 1. Map of recent developments in artificial intelligence ethics worldwide

Data sources: Pan-Canadian AI Strategy [5], Australian Federal Budget 2018-2019 [6] German Ministry of Transport and Digital Infrastructure [1], National Institute for Transformation of India [7], The Villani Report [8], Reuters [9], Japanese Society for Artificial Intelligence [10], European Commission [11] UK Parliament [12], Singapore Government [13] China's State Council [14] New York City Hall [2]

An approach based on case studies. This report examines key issues through exploring a series of case studies and trends that have prompted ethical debate in Australia and worldwide (see Figure 2).

	Examples of case studies	Most relevant principles
Data governance and AI	**Identifying de-identified data** In 2016, a dataset that included de-identified health information was uploaded to data.gov.au. It was expected that the data would be a useful tool for medical research and policy development. Unfortunately, it was discovered that in combination with other publicly available information, researchers were able to personally identify individuals from the data source. Quick action was taken to remove the dataset from data.gov.au.	**Privacy protection** **Fairness**
Automated decisions	**Houston teachers fired by automated system** An AI was used by the Houston school district to assess teacher performance and in some cases fire them. There was little transparency regarding the way that the AI was operating. The use of this AI was challenged in court by the teacher's union, as the system was proprietary software and its inner workings were hidden. The case was settled and the district stopped using it [15].	**Fairness** **Transparency and explainability** **Contestability** **Accountability**
Predicting human behaviour	**The COMPAS sentencing tool** COMPAS is a tool used in the US to give recommendations to judges about whether prospective parolee will re-offend. There is extensive debate over the accuracy of the system and whether it is fair to African Americans. Investigations by a non-profit outlet have indicated that incorrect predictions unfairly categorise black Americans as a higher risk. The system is proprietary software [16-19].	**Do no harm** **Regulatory and legal compliance** **Privacy protection** **Fairness** **Transparency and explainability**

Figure 2. Table of key issues examined in chapters, case studies and relevant principles

Data sources: Office of the Australian Information Commissioner [20], US Senate Community Affairs Committee Secretariat [15], ProPublica [16,18,19], Northpointe [17]

Artificial intelligence (AI) holds enormous potential to improve society. While a "general AI" that replicates human intelligence is seen as an unlikely prospect in the coming few decades, there are numerous "narrow AI" technologies which are already incredibly sophisticated at handling specific tasks [3]. Medical AI technologies and autonomous vehicles are just a few high profile examples of AI that have potential to save lives and transform society.

The benefits come with risks. Automated decisions systems can limit issues associated with human bias, but only if due care is focused on the data used by those systems and the ways they assess what is fair or safe. Automated vehicles could save thousands of lives by limiting accidents caused by human error, but as Germany's Transport Ministry has highlighted in its ethics framework for AVs, they require regulation to ensure safety [1].

Existing ethics in context, not reinvented. Philosophers, academics, political leaders and ethicists have spent centuries developing ethical concepts, culminating in the human-rights based framework used in international and Australian law. Australia is a party to seven core human rights agreements which have shaped our laws [21]. An ethics framework for AI is not about rewriting these laws or ethical standards, it is about updating them to ensure that existing laws and ethical principles can be applied in the context of new AI technologies.

Core principles for AI

1. Generates net-benefits. The AI system must generate benefits for people that are greater than the costs.

2. Do no harm. Civilian AI systems must not be designed to harm or deceive people and should be implemented in ways that minimise any negative outcomes.

3. Regulatory and legal compliance. The AI system must comply with all relevant international, Australian Local, State/Territory and Federal government obligations, regulations and laws.

4. Privacy protection. Any system, including AI systems, must ensure people’s private data is protected and kept confidential plus prevent data breaches which could cause reputational, psychological, financial, professional or other types of harm.

5. Fairness. The development or use of the AI system must not result in unfair discrimination against individuals, communities or groups. This requires particular attention to ensure the “training data” is free from bias or characteristics which may cause the algorithm to behave unfairly.

6. Transparency & Explainability. People must be informed when an algorithm is being used that impacts them and they should be provided with information about what information the algorithm uses to make decisions.

7. Contestability. When an algorithm impacts a person there must be an efficient process to allow that person to challenge the use or output of the algorithm.

8. Accountability. People and organisations responsible for the creation and implementation of AI algorithms should be identifiable and accountable for the impacts of that algorithm, even if the impacts are unintended.

Data is at the core of AI. The recent advances in key AI capabilities such as deep learning have been made possible by vast troves of data. This data has to be collected and used, which means issues related to AI are closely intertwined with those that relate to privacy and data. The nature of the data used also shapes the results of any decision or prediction made by an AI, opening the door to discrimination when inappropriate or inaccurate datasets are used. There are also key requirements of Australia’s Privacy Act which will be difficult to navigate in the AI age [22].

Predictions about people have added ethical layers. Around the world, AI is making all kinds of predictions about people, ranging from potential health issues through to the probability that they will end up re-appearing in court [16]. When it comes to medicine, this can provide enormous benefits for healthcare. When it comes to human behaviour, however, it’s a challenging philosophical question with a wide range of viewpoints [23]. There are benefits, to be sure, but risks as well in creating self-fulfilling prophecies [24]. The heart of big data is all about risk and probabilities, which humans struggle to accurately assess.

AI for a fairer go. Australia’s colloquial motto is a “fair go” for all. Ensuring fairness across the many different groups in Australian society will be challenging, but this cuts right to the heart of ethical AI. There are different ideas of what a “fair go” means. Algorithms can’t necessarily treat every person exactly the same either; they should operate according to similar principles in similar situations. But while like goes with like, justice sometimes demands that different situations be treated differently. When developers need to codify fairness into AI algorithms, there are various challenges in managing often inevitable trade-offs and sometimes there's no "right" choice because what is considered optimal may be disputed. When the stakes are high, it's imperative to have a human decision-maker accountable for automated decisions—Australian laws already mandate it to a degree in some circumstances [25].

Transparency is key, but not a panacea. Transparency and AI is a complex issue. The ultimate goal of transparency measures are to achieve accountability, but the inner workings of some AI technologies defy easy explanation. Even in these cases, it is still possible to keep the developers and users of algorithms accountable [26]. An analogy can be drawn with people: an explanation of brain chemistry when making a decision doesn't necessarily help you understand how that decision was made—an explanation of that person's priorities is much more helpful. There are also complex issues relating to commercial secrecy as well as the fact that making the inner workings of AI open to the public would leave them susceptible to being gamed [26].

Black boxes pose risks. On the other hand, AI "black boxes" in which the inner workings of an AI are shrouded in secrecy are not acceptable when public interest is at stake. Pathways forward involve a variety of measures for different situations, ranging from explainable AI technologies [27], testing, regulation that requires transparency in the key priorities and fairness measures used in an AI system, through to measures enabling external review and monitoring [26]. People should always be aware when a decision that affects them has been made by an AI, as difficulties with automated decisions by government departments have already been before Australian courts [28].

Justifying decisions. The transparency debate is one component feeding into another debate: justifiability. Can the designers of a machine justify what their AI is doing? How do we know what it is doing? An independent, normative framework can serve to inform the development of AI, as well as justify or revise the decisions made by AI. This document is part of that conversation.

Privacy measures need to keep up with new AI capabilities. For decades, society has had rules about how fingerprints are collected and used. With new AI-enabled facial recognition, gait and iris scanning technologies, biometric information goes well beyond fingerprints in many respects [29]. Incidents like the Cambridge Analytica scandal demonstrate how far-reaching privacy breaches can be in the modern age, and AI technologies have the potential to impact this in significant ways. We may need to further explore what privacy means in a digital world.

Keeping the bigger picture in focus. Discussions on the ethics of autonomous vehicles tend to focus on issues like the "trolley problem" where the vehicle is given a choice of who to save in a life-or-death situation. Swerve to the right and hit an elderly person, stay straight and hit a child, or swerve to the left and kill the passengers? These are important questions worth examining [30], but if widespread adoption of autonomous vehicles can improve safety and cut down on the hundreds of lives lost on Australian roads every year, then there is a risk that lives could be lost if relatively far-fetched scenarios dominate the discussion and delay testing and implementation. The values programmed into autonomous vehicles are important, though they need to be considered alongside potential costs of inaction.

AI will reduce the need for some skills and increase the demand for others Disruption in the job market is a constant. However, AI may fuel the pace of change. There will be challenges in ensuring equality of opportunity and inclusiveness [31]. An ethical approach to AI development requires helping people who are negatively impacted by automation transition their careers. This could involve training, reskilling and new career pathways. Improved information on risks and opportunities can help workers take proactive action. Incentives can be used to encourage the right type of training at the right times. Overall, acting early improves the chances of avoiding job-loss or ongoing unemployment.

AI can help with intractable problems. Long-standing health and environmental issues are in need of novel solutions, and AI may be able to help. Australia’s vast natural environment is in need of new tools to aid in its preservation, some of which are already being implemented [32]. People with serious disabilities or health problems are able to participate more in society thanks to AI-enabled technologies [33].

International coordination is crucial. Developing standards for electrical and industrial products required international coordination to make devices safe and functional across borders [34]. Many AI technologies used in Australia won’t be made here. There are already plenty of off-the-shelf foreign AI products being used [35]. Regulations can induce foreign developers to work to Australian standards to a point, but there are limits. International coordination with partners overseas, including the International Standards Organisation (ISO), will be necessary to ensure AI products and software meet the required standards.

Implementing ethical AI. AI is a broad set of technologies with a range of legal and ethical implications. There is no one-size-fits all solution to these emerging issues. There are, however, tools which can be used to assess risk and ensure compliance and oversight. The most appropriate tools can be selected for each individual circumstance.

A toolkit for ethical AI

1. Impact Assessments: Auditable assessments of the potential direct and indirect impacts of AI, which address the potential negative impacts on individuals, communities and groups, along with mitigation procedures.

2. Internal or external review: The use of specialised professionals or groups to review the AI and/or use of AI systems to ensure that they adhere to ethical principles and Australian policies and legislation.

3. Risk Assessments: The use of risk assessments to classify the level of risk associated with the development and/or use of AI.

4. Best Practice Guidelines: The development of accessible cross industry best practice principles to help guide developers and AI users on gold standard practices.

5. Industry standards: The provision of educational guides, training programs and potentially certification to help implement ethical standards in AI use and development

6. Collaboration: Programs that promote and incentivise collaboration between industry and academia in the development of ‘ethical by design’ AI, along with demographic diversity in AI development.

7. Mechanisms for monitoring and improvement: Regular monitoring of AI for accuracy, fairness and suitability for the task at hand. This should also involve consideration of whether the original goals of the algorithm are still relevant.

8. Recourse mechanisms: Avenues for appeal when an automated decision or the use of an algorithm negatively affects a member of the public.

9. Consultation: The use of public or specialist consultation to give the opportunity for the ethical issues of an AI to be discussed by key stakeholders.

Best practice based on ethical principles. The development of best practice guidelines can help industry and society achieve better outcomes. This requires the identification of values, ethical principles and concepts that can serve as their basis.

About this report. This report covers civilian applications of AI. Military applications are out of scope. This report also acknowledges research into AI ethics occurring as part of a project by the Australian Human Rights Commission [36], as well as work being undertaken by the recently established Gradient Institute. This work complements research being conducted by the Australian Council of Learned Academies (ACOLA) and builds upon the Robotics Roadmap for Australia by the Australian Centre for Robotic Vision. From a research perspective, this framework sits alongside existing standards, such as the National Health and Medical Research Council (NHMRC) Australian Code for the Responsible Conduct of Research and the NHMRC's National Statement on Ethical Conduct in Human Research.

A guide to this framework

In this evolving domain, there may be no single ethical framework to guide all decision making and implementation of Artificial Intelligence. The chapters of this ethics framework provide a strong foundation for both awareness and achievement of better ethical outcomes from AI. AI is a broad family of technologies which requires careful, specialised approaches. These chapters provide a broad understanding of AI and ethics, which can be used to identify and begin crafting those specialised approaches. This ethical framework should not be used in isolation from key business or policy decisions, and will supplement fit-for-purpose applications.

Chapter 2: Existing frameworks, principles and guidelines on AI ethics

This chapter identifies and summarises some of the key approaches to issues related to AI and ethics around the world. It helps provide broader context for the current state of AI ethics and highlights strategies that can be observed for lessons on implementation and effectiveness.

Chapter 3: Data governance

This section highlights the crucial role of data in most modern AI applications. It explores the ways in which the input data can affect the output of the AI systems, as well as the ways in which data breaches, consent issues and bias can affect the outcomes derived from AI technologies.

- Data governance is crucial to ethical AI; organisations developing AI technologies need to ensure they have strong data governance foundations or their AI applications risk being fed with inappropriate data and breaching privacy and/or discrimination laws.
- AI offers new capabilities, but these new capabilities also have the potential to breach privacy regulations in new ways. If an AI can identify anonymised data, for example, this has repercussions for what data organisations can safely use.
- Organisations should constantly build on their existing data governance regimes by considering new AI-enabled capabilities and ensuring their data governance system remains relevant.

Chapter 4: Automated decisions

This chapter highlights the ethical issues associated with delegating responsibility for decisions to machines.

- Existing legislation suggests that for government departments, automated decisions are suitable when there is a large volume of decisions to be made, based on relatively uniform, uncontested criteria. When discretion and exceptions are required, automated decision systems are best used only as a tool to assist human decision makers—or not used at all. These requirements are not mandated for other organisations, but are a wise approach to consider.
- Consider human-in-the-loop (HITL) principles during the design phase of automated decisions systems, and ensure sufficient human resources are available to handle the likely amount of inquiries.
- There must be a clear chain of accountability for the decisions made by an automated system. Ask: Who is responsible for the decisions made by the system?

Chapter 5: Predicting human behaviour

This chapter examines the ethical difficulties that emerge when creating systems that are designed to take input data from humans and make judgements about those people.

- AI is not driven by human bias but it is programmed by humans. It can be susceptible to the biases of its programmers, or can end up making flawed judgments based on flawed information. Even when the information is not flawed, if the priorities of the system are not aligned with expectations of fairness, then the system can deliver negative outcomes.
- Justice means that like situations should deliver like outcomes, but different situations can deliver different outcomes. This means that developers need to pay special care to vulnerable, disadvantaged or protected groups when programming AI.
- Full transparency is sometimes impossible, or undesirable (consider privacy breaches). But there are always ways to achieve a degree of transparency. Take neural nets, for example: they are too complex to explain, and very few people would have the expertise to understand anyway. However, the input data can be explained, the outcomes from the system can be monitored, and the impacts of the system can be reviewed internally or externally. Consider the system, and design a suitable framework for keeping it transparent and accountable. This is necessary for ensuring the system is operating fairly, in line with Australian norms and values.

Chapter 6: Current examples of AI in practice

This chapter examines two areas where AI technologies are having a significant impact at this point in time—autonomous vehicles and surveillance technologies.

- Autonomous vehicles require hands-on safety governance and management from authorities, because there are competing visions of how they should prioritise human life and a system without a cohesive set of rules is likely to deliver worse outcomes that are not optimised for Australian road rules or conditions.
- AI-enabled surveillance technologies should consider "non-instrumentalism" as a key principle—does this technology treat human beings as one more cog in service of a goal, or is the goal to serve the best interests of human beings?
- In many ways, biometric data is replacing fingerprints as a key tool for identification. The ease at which AI-enabled voice, face and gait recognition systems can identify people poses an enormous risk to privacy.

Contents

Figures

1 Introduction

> "The machine is only a tool after all, which can help humanity progress faster by taking some of the burdens of calculations and interpretations off its back. The task of the human brain remains what it has always been; that of discovering new data to be analyzed, and of devising new concepts to be tested."
>
> I, Robot, Isaac Asimov

Throughout the 1940s and 1950s, science fiction writer Isaac Asimov published fictional tales of intelligent robots and envisioned three rules to govern them. He would later add a fourth law to protect humanity more broadly. Then and now, it was clear that four rules would be insufficient to handle the philosophical and technical complexity of the task. Asimov's laws pre-date decades of studies into the ethics of artificial intelligence, which arguably began in 1955 when the term artificial intelligence (AI) was coined by mathematician John McCarthy and his colleagues [37]. Today, AI ethics remains a rich and highly relevant field of inquiry.

In this report AI is defined as:

> A collection of interrelated technologies used to solve problems autonomously and perform tasks to achieve defined objectives without explicit guidance from a human being.

Today's AI has capabilities for unaided machine learning and complex problem solving delivered by virtual (e.g. automated online search tools, computerised game simulators) and mechanical systems (e.g. robots, autonomous vehicles). This definition of AI encompasses both recent, powerful advances in AI such as neural nets and deep learning, as well as less sophisticated but still important applications with significant impacts on people, such as automated decision systems.

This report deals exclusively with civilian applications of AI and does not delve into the ethics of AI in the military. This document focuses on "narrow AI" which performs a specific function, rather than "general AI" which is comparable to human intelligence across a range of fields and is not seen as a likely prospect by 2030.

Enormous benefits are already accompanying the age of AI. New AI-enabled medical technologies have the potential to save lives. There are persuasive indications that autonomous vehicles may cut down on the road toll. New jobs are being created, economies are being rejuvenated, and creative new forms of entertainment are emerging.

But some of these tools are powerful and very complex. That means that their design and use are both subject to significant ethical considerations. The report, 'Ethical by design: principles for good technology', by the Ethics Centre in Sydney, provides an overview of the philosophical basis of why an ethical approach to technology matters [38]. It highlights the importance of coming to an "ethical equilibrium" that satisfies a broad range of attitudes toward what is ethical and what is not [38]. Although this AI Ethics Discussion Paper was developed in keeping with this concept, there are a few foundational assumptions that lie at the heart of the document—that we do have power to alter the outcomes we get from technology, and that technology should serve the best interests of human beings and be aligned with human values.

The notion that technology is value-neutral while people make all the decisions is a flawed one. As historian Melvin Kranzberg once said, "technology is neither good nor bad, nor is it neutral [39]." Technology shapes people just as people shape technology. Today, cities have been transformed by road infrastructure to serve cars. Smartphones change our attention spans and have evolved our workforce. Medical technologies such as IVF have even changed the ways children can be conceived. People were born into this transformed world, and it affected the ways they lived their lives. Not everyone gets access to the most advanced technologies, and not everybody gets a say in how they are used once they are released into the public domain. This makes it all the more important to track and consider the implications of new technologies at the time they are emerging. If we accept that we have the ability to determine the outcomes we get from AI, then there is an ethical imperative to try to find the best possible outcomes and avoid the worst.

Around the world, people are being given prison sentences based on assessments from autonomous systems. The world of transportation faces a possible wave of disruption as automated vehicles move on to the roads, displacing jobs and creating new ones. AI is watching people through surveillance, sometimes improving safety, sometimes encroaching on privacy. People are being assessed by AI for likely medical problems, while others are being assessed to gauge their consumer preferences.

The effects of AI will be transformative for Australian society. Countries everywhere are developing plans for an AI-enabled era. In the past two years the United States, China, the United Kingdom, India, Finland, Germany, the European Commission and other countries and organisations have published AI strategies [40]. An important component of these national strategies is the ethical issues raised by the advancement and adoption of AI technologies.

This ethics framework highlights the ethical issues that are emerging or likely to emerge in Australia from AI technologies and outlines the initial steps toward mitigating them. It does not reinvent ethical concepts, but contextualises existing ethical considerations developed over centuries of practice in order to keep pace with the new capabilities that are emerging via AI. It seeks pragmatic solutions and future pathways in this rapidly evolving area by analysing case studies, while acknowledging the importance of ongoing theoretical and philosophical discussions of the implications of AI technology.

The development and adoption of advanced forms of narrow AI will not wait for government or society to catch up—these technologies are already here and developing quickly. Blocking all of these technologies is not an option, any more than cutting off access to the internet would be, but there may be scope to ban particularly harmful technologies if they emerge. As with the internet, there are risks involved in the use of AI, but they should not be seen as a reason to reject it entirely. Many AI-driven technologies have been proven to save lives and reduce human suffering, thus, an ethical approach to AI is not a restrictive one. There have already been cases where the slow pace of regulatory adaptation has hindered the development of potentially life-saving AI technologies [41]. Numerous stakeholders consulted during the formulation of this report expressed the concern that over-regulating this space could have negative consequences and drive innovation offshore, to the detriment of smaller Australian companies and to the advantage of established multinationals with more resources.

With that in mind, it is also important to consider the consequences of taking no action in steering the ethical development and use of AI in Australia. As the case studies in this document demonstrate, AI technologies are already having a range of effects on people around the world. The developers of these technologies are working in an area that is not yet well regulated, which means they are exposed to added risk. If any backlash occurs, they run the risk of making mistakes or being scapegoated for problems which could potentially be avoided if the area was well understood and proper rules, regulations or ethical guidance were in place.

This report emphasises real world case studies specifically related to AI and automated systems, rather than a detailed exploration of the philosophical implications of AI, but those philosophical inquiries are also important. The goal of this document is to provide a pragmatic assessment of key issues to help foster ethical AI development in Australia. It has been written with the goal of creating a toolkit of practical and implementable methods (such as developing best practice guidelines or providing education and training) that can be used to support core ethical principles designed to assist both AI developers and Australia as a whole. Further research and analysis by professional ethicists will be necessary as AI technologies continue to shape Australian society.

This Ethics Framework provides guidance on how to approach the ethical issues that emerge from the use of AI. This report argues that AI has the potential to provide many social, economic, and environmental benefits, but there are also risks and ethical concerns regarding privacy, transparency, data security, accountability, and equity.

An ethical framework such as this is one part of suite of governance mechanisms and policy tools which can include laws, regulations, standards and codes of conduct. An ethical framework on its own will not ensure the safe and ethical development and use of AI. Fit for purpose, flexible and nimble approaches are appropriate for the regulation and governance of new and emerging digital technologies. Ethics both inform and are informed by laws and community values. These principles take laws into account and can form the groundwork for the formulation of more specific codes, laws or regulation, but are intended as a guide only.

In developing and governing AI technologies, neither over-regulation nor a laissez-faire approach is sufficient. There is a path forward which allows for flexible solutions, the fostering of innovation and a firm dedication to aligning the development of AI with human values.

This document does not aim to provide legal guidance. Regulations and possibly legal reform should be formulated as needed by the appropriate legal and governing bodies, for each specific domain or application. The goal of this document is to help identify ethical principles and to elicit discussion and reflection of how AI should be developed and used in Australia.

With a proactive approach to the ethical development of AI, Australia can do more than just mitigate against risks—if we can build AI for a fairer go, we can secure a competitive advantage as well as safeguard the rights of Australians.

2 Existing frameworks, principles and guidelines on AI Ethics

The following documents and publications provide an outline of relevant legislation and ethical principles relating to the use and development of AI. The literature is sourced from governments and multilateral organisations both within Australia and internationally. This summary is not a systematic review of all available literature relating to the ethical use of AI, but a collection of key documents that give a high-level overview of the current state of AI ethics. They have been selected on the basis of impact and visibility.

2.1 Australian frameworks

Artificial intelligence is a broad set of technologies with applications across virtually all industries and aspects of government and society. Government agencies are already using automated decisions systems to streamline the provision of services, and there is existing advice that provides some insight on governance and oversight of AI.

2.1.1 Government and automated decisions

Some key documents authored by government bodies provide background on how agencies should use AI. This includes section 6A of the Social Security (Administration) Act 1999, which states:

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 [25].

This is just one of numerous legislative clauses allowing government agencies to use computers for decision-making – since 2010, the departments of Social Services, Health, Education and Training, Immigration and Border Protection, Agriculture and Water Resources and Veterans' Affairs have all been given some authority to let automated systems make decisions [42]. This law clarifies an important aspect of AI ethics as expressed in Australian legislation: when decisions are made by automated systems, a human being with authority must be accountable for those decisions.

In 2003, a Department of Finance working group for Automated Assistance in Administrative Decision Making released a best practice guide for government agencies seeking to use AI to make decisions [43]. The guide, updated in 2007, outlines 27 principles covering a range of issues, from review mechanisms through to the appropriate ways to override a decision made by an automated system. The guidelines include flow charts of how automated decisions should be made, and checklists to help ensure that automated decisions are being made according to the values of administrative law. These checklists can help serve as a valuable starting point for developing toolkits for AI use in administration.

The guide distinguishes between two key types of decisions: administrative decisions for which the decision-maker is required to exercise discretion; and those for which no discretion is exercisable once the facts are established. Given the high volume of routine decisions that need to be made by some agencies, the guide judged it suitable to use automated systems in making decisions where no discretion was required. In other cases, automated decision-making systems were determined to be best used as 'decision-making tools' for human supervisors. This distinction clarifies that while AI can be a valuable tool in decision-making, there are some decisions requiring human judgment, particularly in the context of public policy administration.

Federal government agencies are also developing AI-specific practices. An interdepartmental committee on AI regularly convenes to discuss how government agencies can utilise AI. As automation becomes more pervasive within government, industry, and broader society, frameworks such as the best practice guide on automated decision-making can help to ensure that government bodies remain accountable to the public.

Guidance may also be sought from other examples of government action around automated decision-making. For instance, the New York City government is the first American government body to set up a task force specifically to examine accountability in automated decisions. The Automated Decisions Task Force will examine automated systems through the lens of equity, fairness and accountability, and is set to release a report in December 2019 that will recommend procedures for reviewing and assessing algorithmic tools used by the city [2].

2.1.2 Australia's international human rights obligations and anti-discrimination legislation

Australia is a signatory to seven core international human rights agreements [21]

- The International Covenant on Civil and Political Rights (ICCPR) [44]
- The International Covenant on Economic, Social and Cultural Rights (ICESCR) [45]
- The International Convention on the Elimination of All Forms of Racial Discrimination (CERD) [46]
- The Convention on the Elimination of All Forms of Discrimination against Women (CEDAW) [47]
- The Convention against Torture and Other Cruel, Inhuman or Degrading Treatment or Punishment (CAT) [48]
- The Convention on the Rights of the Child (CRC) [49]
- The Convention on the Rights of Persons with Disabilities (CRPD) [50]
- These agreements are all derived from the Universal Declaration of Human Rights which was released in 1948 [36,51]

Australia is also a party to a number of related protocols.

Under Australia's Human Rights (Parliamentary Scrutiny) Act 2011, new bills must be accompanied by a statement of compatibility that demonstrates how they align with the seven aforementioned human rights agreements [52,53]. The Parliamentary Joint Committee on Human Rights scrutinises laws to confirm they are compatible with Australia's human rights obligations [52]. Any future Australian legislation will need to abide by these principles amid change occurring due to AI.

The Australian Human Rights Commission is currently in the process of developing a report examining Australia's human rights obligations in the context of emerging technological issues. The report will be released in 2020 after public consultation, but an issues paper has already been released for discussion [36].

In addition, Australia has a number of anti-discrimination laws at both state and federal levels. Federal laws include the Age Discrimination Act 2004, the Disability Discrimination Act of 1992, the Racial Discrimination Act of 1975 and the Sex Discrimination Act of 1984 [54]. Measures to combat discrimination are highly relevant to AI, as AI systems are vulnerable to discriminatory outcomes – for instance, there have been cases where AI systems have used historical data, leading to results that replicated the biases or prejudices of that original data, as well as any flaws in the collection of that data [55]. In ensuring that AI systems and programs are created in accordance with existing anti-discrimination laws, designers will need to consider the likely outcomes caused by their algorithms during the design phase.

2.1.3 Data-sharing legislation in Australia

Data is a key component of AI. It is necessary for both developing the skills needed to work on AI and the technology itself, as large datasets are often required to 'teach' machine learning technologies. Legislation that guides data-sharing therefore affects the development of AI, but is also highly relevant to the privacy of all Australians.

A key document on data-sharing in Australia is a 2017 report from the Productivity Commission, *Data Availability and Use* [56]. The report focuses on ways to streamline access to data, as well as exploring the economic benefits that could be gained through improved data access. The report covers several areas of particular relevance to this ethics framework, including:

- Assisting individuals to access their personal data being held by public agencies
- Identifying datasets with high value to the public
- The role of third-party intermediaries in assisting consumers to make use of their data
- The benefits and costs of data standardisation and public releases (which has relevance for the broader development of AI and how personal information may be handled by AI systems)

As a part of the Australian Government's data reform efforts, a Data Sharing and Release bill is being formulated. The Department of Prime Minister and Cabinet has released a discussion paper outlining some key principles of the bill, including the following goals [57]:

- To safeguard data sharing and release in a consistent and appropriate way
- To enhance the integrity of the data system
- To build trust in use of public data
- To establish institutional arrangements for data governance, via a National Data Commissioner and its supporting office
- To promote better sharing of public sector data

The Office of the Victorian Information Commissioner has also released an issues paper outlining key questions relating to data used in AI [58]. The report is particularly concerned with exploring potential privacy issues arising from the development and use of AI. It promotes the use of 'ethical data stewardship', which requires a commitment to transparency and accountability in the way data is collected and used. The report also proposes the need for independent governance and oversight of the AI industry, to ensure that the principles of ethical data stewardship are adhered to.

Data-sharing practices are an integral aspect of AI ethics. AI systems require effective facilitation of data-sharing and collection in order to function and develop – however, it is crucial that this process does not compromise privacy. Comprehensively reviewing and reforming Australia's data-sharing practices in order to strike this balance would help resolve some key ethical issues associated with AI development, by reducing the possibility that AI programs could access and misuse personal information.

2.1.4 Privacy Act

Privacy issues associated with the internet are not new but AI has the potential to amplify existing challenges. The Australian Privacy Act 1988 (Privacy Act) regulates how personal information is handled. The Privacy Act defines personal information as [59]:

…information or an opinion, whether true or not, and whether recorded in a material form or not, about an identified individual, or an individual who is reasonably identifiable.

Common examples are an individual's name, signature, address, telephone number, date of birth, medical records, bank account details and commentary or opinion about a person.

The Privacy Act includes thirteen Australian Privacy Principles (APPs) [60] , which apply to some private sector organisations, as well as most Australian and Norfolk Island Government agencies. These are collectively referred to as 'APP entities'. The Privacy Act also regulates the privacy component of the consumer credit reporting system [61], tax file numbers [62], and health and medical research [63].

The Office of the Australian Information Commissioner is responsible for privacy functions that are conferred by the Privacy Act.

2.2 International frameworks

Many of the AI strategies developed by governments around the world include a discussion of ethics, and this information is important in framing the international context for Australia's approach. In particular, key ethical questions are explored in the national strategies of the United Kingdom, France and Germany, all of which have been shaped by the European Union's data protection laws.

In 2018, the EU began implementing its General Data Protection Regulation (GDPR) which is among the largest, most far-reaching data-sharing laws in the world. It includes the 'right to be forgotten', which requires organisations with data operations in the EU to have measures in place allowing members of the public to request the removal of personal information held on them. Another element of the GDPR is 'privacy by design', which clarifies statutory requirements for privacy at the system design phase, The GDPR also encourages (but does not enforce) certification systems. The GDPR also includes sections relevant to automated decisions, indicating that automated decisions systems cannot be the sole decision-making entity when the decision has legal ramifications. Article 22 states: "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 [64]."

Academics have pointed out that the language of this article is vague and that a right to explanations from automated system may not actually exist under the GDPR [65].

The European Union also has an official plan for AI development – *Artificial Intelligence for Europe* – which explicitly highlights the digital single market as a key driver of AI development, and emphasises the creation of ethical AI as a competitive advantage for European nations [4]. In a 2018 statement, the European Group on Ethics in Science and New Technologies suggested that a global standard of fundamental principles for ethical AI, supported by legislative action, is required to ensure the safe and sustainable development of AI [66].

The European Commission has also issued *Draft Ethics Guidelines for Trustworthy AI*, which emphasise that AI should be "human centric" and "trustworthy [67]." These two points emphasise not only the ethics of AI, but also certain technical aspects of AI, because "a lack of technological mastery can cause unintentional harm." It outlines a framework for trustworthy AI that begins with an "ethical purpose" for the AI, then moves to the realisation of that AI, followed by requirements and finally technical and non-technical methods of oversight [67].

The United Kingdom's national plan for AI (*AI in the UK, Ready, Willing and Able?)* explores AI ethics from numerous angles, with sections on inequality, social cohesion, prejudice, data monopolies, criminal misuse of data, and suggestions for the development of an AI Code. The report points out that there are numerous state and non-state actors developing ethical principles for the use of AI, but a coordinated approach is lacking in many cases. According to the report, "mechanisms must be found to ensure the current trend for ethical principles does not simply translate into a meaningless box-ticking exercise." [3] The report also nominates the Alan Turing Institute as the national centre for AI research, with part of its mandate being further exploration of the ethics of artificial intelligence. The document includes a code with five key elements:

1. Artificial intelligence should be developed for the common good and benefit of humanity.
2. Artificial intelligence should operate on principles of intelligibility and fairness.
3. Artificial intelligence should not be used to diminish the data rights or privacy of individuals, families or communities.
4. All citizens have the right to be educated to enable them to flourish mentally, emotionally and economically alongside artificial intelligence.
5. The autonomous power to hurt, destroy or deceive human beings should never be vested in artificial intelligence. [3]

The French national report on AI examines a number of key ethical issues and proposes measures to address these [68]. For instance, 'discrimination impact assessments' are suggested as one possible measure to address hidden bias and discrimination in AI, citing the existence of 'privacy impact assessments' in European law. The report also explores the 'black box problem'—it is easy to explain the data going in to the AI program and easy to explain the data that comes out, but what occurs within is difficult for most people to understand. As such, technologies that 'explain' AI processes will be increasingly important as AI becomes more commonly used. The report also extensively canvasses the issue of automation, and the need for retraining measures to mitigate its impact on the workforce. At a regulatory level, the report emphasises that designing procedures, tools, and methods that allow for the auditing of AI systems will be key in ensuring that the systems conform to legal and ethical frameworks. It also suggests that it will be necessary to "instate a national advisory committee on ethics for digital technology and artificial intelligence, within an institutional framework [68]."

In Germany, the national report *Automated and Connected Driving* is the world's most comprehensive ethics report into autonomous vehicles (AVs) to date. The report lays out key principles for the development of AVs, explicitly stating that the public sector is responsible for safety and that licencing of automated systems is a key requirement. The report emphasises that while the personal freedom of the individual is a paramount concern of government, this must be pursued within the context of public safety. The prioritisation of human life is a key element of this ethical framework – where damage is inevitable, animals or property should never be placed above human life. When human life must be damaged, the German ethics framework states that: "any distinction based on personal features (age, gender, physical or mental constitution) is strictly prohibited. However, general programming to reduce the number of personal injuries may be justifiable" [1] . The report also notes that ethical 'dilemma situations' depend on the actual specific situation and cannot be standardised or programmed – as such, it would be desirable for an independent public sector agency to systematically process the lessons learned from these situations.

However, it may still prove necessary to program vehicles to deal with these ethical dilemma situations, which would indicate some degree of standardisation. While humans are not expected to be able to make well-reasoned decisions in the brief moment before an accident, this may not be the case for autonomous vehicles which can act rapidly but require programming beforehand. "The court understands that if you've only been given one second to make a decision, you might make a decision that another reasonable person might not have made," Dr Finkel told media [69]. "Will we be as generous to a computerised algorithm that can run at much faster speeds than we can? I don't know."

2.3 Organisational and institutional frameworks

2.3.1 Australian Council of Learned Academies

The Australian Council of Learned Academies (ACOLA) is compiling a comprehensive horizon scan of issues affecting the development of AI in Australia. It identifies social impacts of AI that will affect Australia and New Zealand, with input from key academics in the field of AI. The report covers the relevance of AI to key industries like agriculture, fintech, and transport, as well as the ways in which AI affects government and social policy.

This report is being prepared concurrently with the ACOLA report. Of particular relevance to this ethical framework are discussions of individual agency and autonomy, and of how AI can affect an individual's sense of self. Other elements of the report cover social licence, inclusion, privacy and data bias in AI, as well as the differing concepts of fairness in algorithms. The ACOLA report should be considered complementary to this framework, and when released will provide additional analysis that can help policymakers understand key issues relating to AI.

2.3.2 Nuffield Foundation's roadmap for AI research

Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research by the Nuffield Foundation, examines the ethical implications of research into AI. It first examines the ambiguity in many of the key concepts that are regularly brought up in discussions of AI ethics, such as values and privacy, which can hold different meanings among different audiences. It aims to ensure that when discussing these issues, people do not "talk past one another". It also makes the key point that a number of values are often in conflict with each other and there will inevitably be tradeoffs—for example, quality of services can often be in conflict with privacy; convenience can be in conflict with dignity and accuracy can be in conflict with fairness [70]. The inevitability of tradeoffs in AI algorithms is discussed further in chapter 5.3 of this report.

2.3.3 Institute of Electrical and Electronics Engineers

Some of the most comprehensive documents regarding the ethical development of AI have been produced by the Institute of Electrical and Electronics Engineers (IEEE) through their Global Initiative on Ethics of Autonomous and Intelligent Systems, which is comprised of several hundred world leaders across industry, academia, and government [71]. In 2016 the group released an initial report on ethical design, [72] and based on public feedback released the second version for review in 2017 [73]. Their primary goal is to produce an accessible and useful framework that can serve as a robust reference for the global development of AI.

The IEEE outlines five core principles to consider in the design and implementation of AI:

- Adherence to existing human rights frameworks
- Improving human wellbeing
- Ensuring accountable and responsible design
- Transparent technology
- Ability to track misuse

The comprehensive and collaborative approach to the development of the framework provides a well-rounded frame of reference for company, governmental and academic ethical guidelines.

2.3.4 AI Now

The US-based *AI Now 2017 Report* [74] reviews current academic research around emerging and topical AI issues. The report focuses on four key issues: labour and automation; bias and inclusion; rights and liberties; and ethics and governance. The discussion of bias and inclusion is the most comprehensive and crucial section of the report, as this issue will impact AI design from the outset and will have long-running negative consequences if not appropriately addressed. According to the report, the current known bias in US-developed AI can be attributed to the lack of gender and ethnic diversity in the tech industry. However, this issue may have global reach, as many tech branches of international companies are based in the US and are thus subject to the same problem. In December of 2018, AI Now issued its *AI Now 2018 Report*, which included a timeline of key ethical breaches involving AI technologies throughout the year [75]. It also highlighted key developments in ethical AI research and emerging strategies to combat bias, such as recognising allocative and representational harms, new observational fairness strategies, anti-classification strategies (which focus on appropriate input data and measuring results), classification parity (equal performance across groups, even at a cost to accuracy among certain groups in some cases) and calibration strategies. The report included significant sections on hidden labour chains in the production of AI technologies. It also highlighted the fact that ethics frameworks on their own are not enough, because concrete actions need to be taken to ensure accountability and justice. AI Now has also produced a template for Algorithmic Impact Assessments, which is discussed further in section 10 of this report [26].

2.3.5 The One Hundred Year Study on Artificial Intelligence

The One Hundred Year Study on Artificial Intelligence, based at Stanford University and launched in 2014, is an effort to detail the long-term influence of AI on society and individuals. A new report is scheduled for release every five years, with the aim of creating a collection of reports that chronicle the development of AI – and the issues raised by that development – over the course of one hundred years [76]. Primarily focussed on North American societies, the report identifies eight areas that will likely undergo the biggest transformation as a result of AI: transport; healthcare; education; low resource communities; public safety; workplaces; homes; and entertainment. Ethical issues associated with each of these areas are highlighted, but the report focuses mainly on the current and future direction of AI in various domains. The authors suggest that restrained government regulation and high levels of transparency around AI development will provide the best climate for encouraging socially beneficial innovation.

2.3.6 The Asilomar AI Principles

In 2017, an AI conference hosted by US organisation the Future of Life Institute reviewed and discussed some of the key literature on AI and developed a list of 23 key principles, known as the Asilomar AI Principles [77]. These have so far garnered 1,273 signatures of agreement from AI researchers and 2,541 signatures from other endorsers. There are 13 principles in the 'ethics and values' section of the report. According to these, the onus is on the AI developer to adhere to responsible design, with the aim of bettering humanity, and AI systems should be designed in line with accepted values and cultural norms, while protecting individual privacy and remaining transparent. Humans should also remain in control of how and whether to delegate decisions to AI systems, with the goal of accomplishing human-chosen objectives [77].

2.3.7 Universal Guidelines for Artificial Intelligence

The Public Voice coalition, a group of NGOs and representatives assembled by the Electronic Privacy Information Center, in October 2018 issued 12 guidelines for the development of AI. These guidelines are based on the premise that "primary responsibility for AI systems must reside with those institutions that fund, develop, and deploy these systems" [78] . The 12 guidelines are:

1) Right to Transparency. All individuals have the right to know the basis of an AI decision that concerns them. This includes access to the factors, the logic, and techniques that produced the outcome.
2) Right to Human Determination. All individuals have the right to a final determination made by a person.
3) Identification Obligation. The institution responsible for an AI system must be made known to the public.
4) Fairness Obligation. Institutions must ensure that AI systems do not reflect unfair bias or make impermissible discriminatory decisions.
5) Assessment and Accountability Obligation. An AI system should be deployed only after an adequate evaluation of its purpose and objectives, its benefits, as well as its risks. Institutions must be responsible for decisions made by an AI system.
6) Accuracy, Reliability, and Validity Obligations. Institutions must ensure the accuracy, reliability, and validity of decisions.
7) Data Quality Obligation. Institutions must establish data provenance, and assure quality and relevance for the data input into algorithms.
8) Public Safety Obligation. Institutions must assess the public safety risks that arise from the deployment of AI systems that direct or control physical devices, and implement safety controls.
9) Cybersecurity Obligation. Institutions must secure AI systems against cybersecurity threats.
10) Prohibition on Secret Profiling. No institution shall establish or maintain a secret profiling system.
11) Prohibition on Unitary Scoring. No national government shall establish or maintain a general-purpose score on its citizens or residents.
12) Termination Obligation. An institution that has established an AI system has an affirmative obligation to terminate the system if human control of the system is no longer possible.

2.3.8 The Partnership on AI

Private companies are increasingly aware of the need for an ethical framework when using and developing AI. The collegiate attitude adopted by traditionally competitive tech companies is an indication of the importance of openness and collaboration when developing said framework. For example, the Partnership on AI, originally established by a handful of large tech companies, is now made up of a wide variety of industry and academic professionals working together to better understand the impacts of AI on society [79]. Rather than a comprehensive ethics framework, the group has outlined eight tenets that their members attempt to uphold. These tenets follow fairly standard topics on the ethical development and use of AI, focusing in particular on technology that benefits as many people as possible; ensuring personal privacies are protected; and encouraging transparency. At this point, the Partnership on AI has not discussed the need to reduce bias and increase diversity in the tech industry.

2.3.9 Google

In June 2018, Google published its company principles in regards to the development of AI [80], after staff within the organisation protested. In addition to the familiar principles regarding safeguarding privacy, developing AI that is beneficial for humanity, and addressing bias, Google has also released a list of AI applications they have chosen not to pursue – including (but not limited to) weapons or other technologies with the principal purpose of causing harm; technologies that gather surveillance in a way that violates internationally accepted norms; and technologies whose purpose contravene principles of international law and human rights.

The response to these principles has not been without scepticism, likely as a result of recent controversies around Google's contracts with the US military (which the company has recently decided not to renew) [81] . Critics also noted that Google had the opportunity to be much more specific and action-oriented in their principles and code, especially as they are touted as being concrete standards actively governing Google's AI research [82,83]. For instance, while the principles stated that Google will seek to avoid bias when developing AI algorithms, a meaningful explanation of how this will be achieved was not addressed. In addition, the proviso for independent review of Google's AI technology development would likely be well received.

Google's subsidiary company Deepmind has also created an ethics board, however it has been criticised for a lack of transparency in both membership and decision-making [84].

2.3.10 Microsoft

Microsoft has also been a prominent voice in the AI ethics debate. In December 2018, Microsoft President Brad Smith wrote on the company's blog that Microsoft believed governments needed to regulate facial recognition and that it was necessary to "ensure that this technology, and the organizations that develop and use it, are governed by the rule of law [85]". The company has also put together a number of principles and tools geared toward ethical AI. Its site includes six key principles: Fairness, inclusiveness, reliability and safety, transparency, privacy and security, and accountability. It has also issued guidelines for responsible bots, which examine how they can earn trust [86].

2.3.11 IBM

As a key player in the computing space and the developer of the question-and-answer AI Watson, IBM has also released a set of materials on AI and ethics. In addition to guidance on ethical AI research and trust and transparency measure, IBM has also released an AI ethics guide for developers. The guide focuses on five key areas for developers: Accountability, Value Alignment, Explainability, Fairness and User Data Rights [87]. It stresses that the ethical development of AI cannot solely be viewed as a "technical" problem to be resolved, and instead requires a strong focus on the communities it affects.

2.3.12 The Future of Humanity Research Institute

The University of Oxford's Future of Humanity Institute calls for research on building frameworks that ensure the socially beneficial development of AI [88]. Their report *AI Governance: A Research Agenda* focuses on developing a global governance system to protect humanity from extreme risks posed by future advanced AI. The report highlights the need for AI leaders to constitutionally commit to developing AI for the common good. While the authors acknowledge that a solution that satisfies the interest of such a diverse range of stakeholders will be exceedingly difficult and complicated, they argue that the potential benefits to society make it a worthy endeavour.

2.3.13 The Ethics of Artificial Intelligence

In the first section of their 2014 publication Bostrom and Yudkowsky discuss the ethical issues associated with machine learning AI developed in the near future [89]. They make the observation that if a machine is going to carry out tasks previously completed by humans then the machine is required to complete the function to the same level that humans do, with "responsibility, transparency, auditability, incorruptibility, predictability, and a tendency to not make innocent victims scream with helpless frustration" [89] . The latter sections of their publication address potential ethical issues associated with super-intelligent machines of the future, but this is out of scope for the current report.

2.3.14 The AI Initiative

Based at Harvard Kennedy School, the AI Initiative has developed a short series of recommendations to help shape global AI policy framework. These are [90]:

- Convene a yearly interdisciplinary meeting to discuss the pressing ethical issues in the development of AI.
- Create a global framework that supports the ethical development of AI, including agreement on beneficial safeguards, transparency standards, design guidelines, and confidence-building measures.
- Implement agreed-upon rules and regulations at local and international levels.

2.4 Key themes

While it is important to note that there exists other relevant work which cannot be reviewed here due to length considerations, the publications discussed provide a snapshot of the current state of AI ethics frameworks, and assist in framing the context of Australia's own uniquely tailored framework. Collectively, the literature emphasise that the principles required for developing ethical AI centre on responsible design that benefits humanity. This benefit is achieved through protecting privacy and human rights, addressing bias, and providing transparency around the workings of machines. A number of tools have been suggested to support the ethical development and use of AI, including impact assessments, audits, consumer data rights, oversight mechanisms, and formal regulation.

3 Data governance

"But the plans were on display … on display in the bottom of a locked filing cabinet stuck in a disused lavatory with a sign on the door saying 'Beware of the Leopard"

Hitchhikers Guide to the Galaxy, Douglas Adams

Issues relating to AI ethics are intertwined with data sharing and use. The age of big data is here and people's opinions, interactions, behaviours and biological processes are being tracked more than ever [91]. However, Australians are largely unaware of the scale and degree to which their data is being collected, sold, shared, collated and used. In one 2018 study, most people surveyed were aware that data generated from their online activities could be tracked, collected and shared by organisations (see Figure 3). However, they frequently reported being unaware of the extent and purpose of for which their data was collected, used and shared [91]. In addition, the study found that Australians were rarely able to grasp the full implications of the terms of use applying to many services such as social media, or products like smartphones [91].

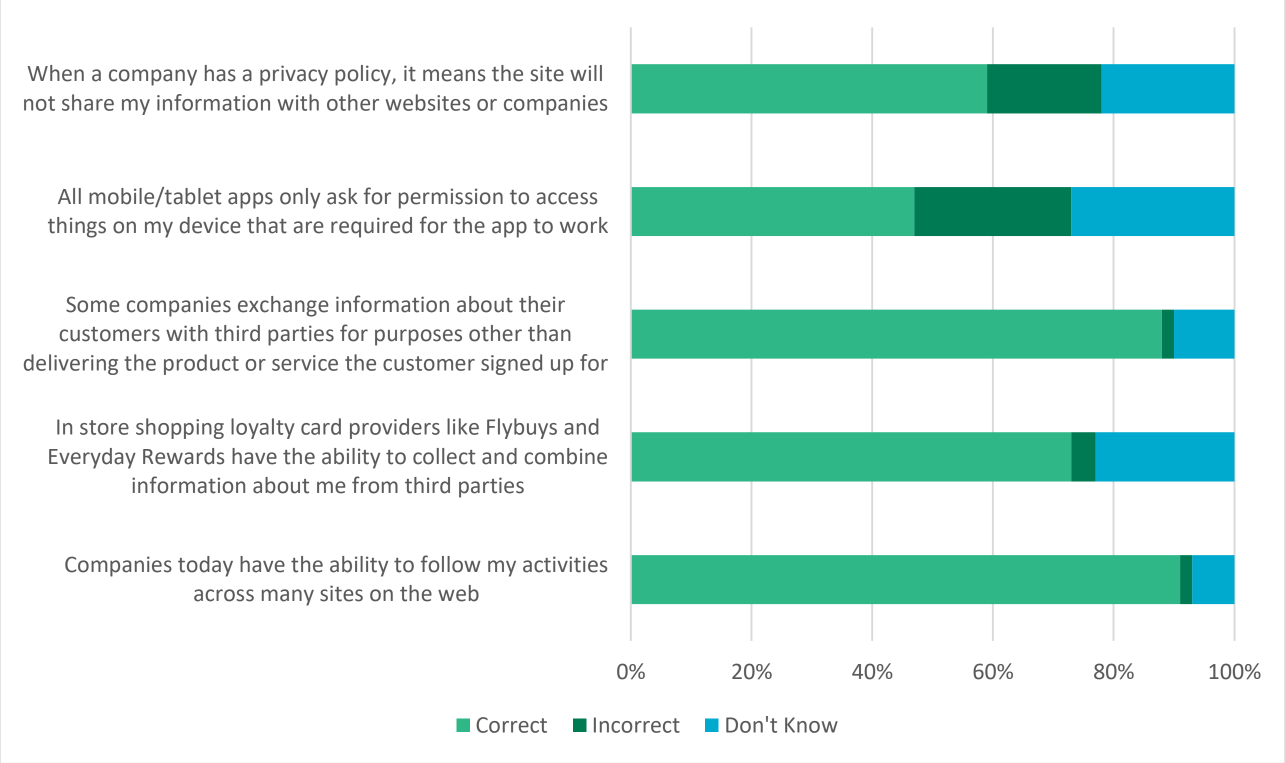


Figure 3. Chart indicating Australian knowledge about consumer data collection and sharing

Data source: Consumer data and the digital economy - Emerging issues in data collection, use and sharing [91]

Despite low levels of public understanding, data governance issues are crucial and will only become more important as AI development gains pace. Data has immense and growing value as the input for AI technologies. As the value and the potential for exploitation of data increases, so does the need to protect the data rights and privacy of Australians.

3.1 Consent and the Privacy Act

Personal data is regulated by the Australian Privacy Act, which classifies it as "information or an opinion, whether true or not, and whether recorded in a material form or not, about an identified individual, or an individual who is reasonably identifiable" [22] .

Privacy itself is a contested term, subject to many varying interpretations. It is far more than a right to a degree of secrecy. Privacy is explicitly stated to be a human right under Article 12 of the Universal Declaration of Human rights.

When working with personal data, protecting the consent process is fundamental to protecting privacy. Due to the sensitive nature of personal data, consent should be adequately addressed at the point of data collection. The Privacy Act stipulates that consent may be express or implied, and that it must abide by four key terms:

- The individual is adequately informed before giving consent
- The individual gives consent voluntarily
- The consent is current and specific
- The individual has the capacity to understand and communicate their consent [22]

The third term of the Privacy Act states that consent must be current, however, at the time of writing there are no specific provisions for the 'right to be forgotten', which features in the EU's recently established General Data Protection Regulation [92] and the UK's updated data protection laws [93]. To align with this international legislation, the 'right to be forgotten' could be considered for future incorporation into Australia's Privacy Act, but there may be other measures that are more suitable in the Australian context. Although this right affords individuals the greatest control over their data, it may be difficult to enforce and adhere to, especially if the data has already been integrated into an AI system and a model has already been trained. It may be instructive to observe how the right to be forgotten is implemented and enforced in the EU, as it is still in the early stages of implementation and review.

3.1.1 Case study: Cambridge Analytica and public trust

The Cambridge Analytica scandal exemplifies the consequences of inadequate consent processes or privacy protection. Through a Facebook app, a Cambridge University researcher was able to gain access to the personal information of not only users who agreed to take the survey, but also the people in those users' Facebook social networks. In this way, the app harvested data from millions of Facebook users. Various reports indicate that these data were then used to develop targeted advertising for various political campaigns run by Cambridge Analytica.

When news broke of this alleged breach in privacy, many felt that Facebook had not provided a transparent consent process. The ability for one user to effectively give consent for the use of others' data was particularly concerning. The allegation that Cambridge Analytica used personal data to profile and target political advertising to the users without appropriate consent was widely criticised [94] and both Cambridge Analytica and Facebook were put under governmental and media scrutiny concerning their data practices. Cambridge Analytica has now become insolvent and Facebook stocks plummeted following the publication of the story (although they recovered their full value eight weeks later) [95,96] .

For industry, this incident serves as an example of the cost of inadequate data protection policies and also demonstrates that it may not be sufficient to merely follow the letter of the law. To avoid repeating these mistakes, consent processes should ensure that consent is current, specific, and transparent. Regular review of data collection and usage policies can help to safeguard against breaches. At a broader level, a balance needs to be struck between protecting individual privacy and ensuring transparent consent processes, while also encouraging investment and innovation in new technologies that require rich datasets.

3.2 Data breaches

With vast amounts of data being collected on individuals, the importance of protecting privacy – and of knowing when privacy has been compromised – is crucial. A recent amendment to the Australian Privacy Act addresses some of these concerns through the Notifiable Data Breaches (NDB) scheme, which stipulates that if personal data is accessed or disclosed in any unauthorised way that may cause harm, all affected individuals must be notified [97].

Between April and June 2018 there were 242 notifications of data breaches in Australia [98]. The majority of those breaches were a result of human error or malicious attacks (see Figure 4), suggesting that there are security gaps in the storage and use of data. Data breaches are costly for organisations, with financial and legal consequences as well as reputational damage. In 2017, the average cost of a data breach in Australia was $2.51 million [99].

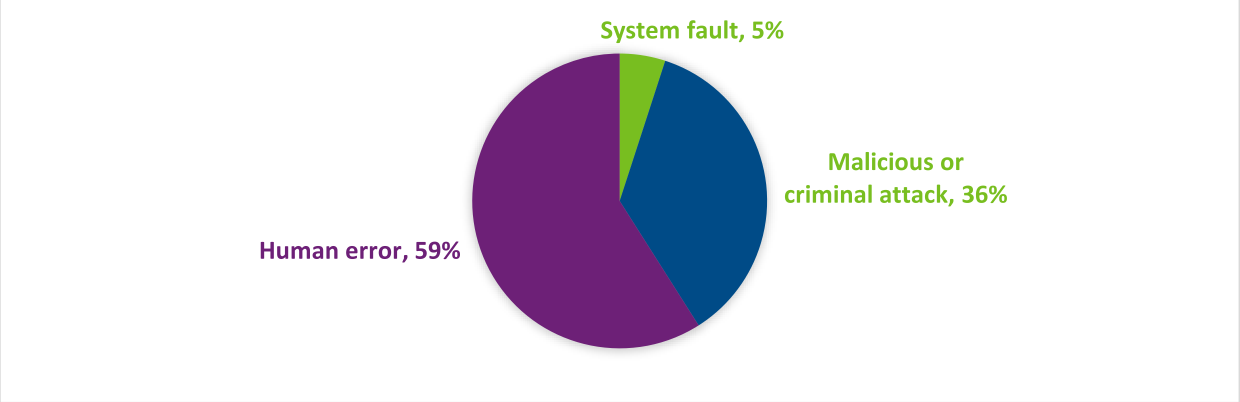


Figure 4. Pie chart showing reasons for Australian data breaches, April-June 2018

Data source: Notifiable Data Breaches Quarterly Report, 1 April-30 June 2018 [98]

The mandatory reporting of breaches under the NDB scheme is a positive move towards ethical data practices in Australia. However, these reforms should be supported with education and training on data protection, as well as regular assessment of data practices to ensure that Australians can trust the security of their private information.

3.2.1 Case study: Equifax data breach

In 2017 Equifax, a US-based credit reporting agency, experienced a data breach affecting at least 145.5 million individuals, with various degrees of sensitive personal information compromised [100]. This breach was particularly concerning as Equifax had the opportunity to prevent the breach – via a patch that had already been available for several months – but failed to identify vulnerabilities and detect attacks to its systems [100]. In addition, due to the huge number of people affected, it took several weeks to identify the individuals and notify the public that the breach had occurred. The cost of the breach was estimated to be in the realm of US$275 million.

It is widely speculated that Equifax did not have appropriate measures or processes in place to adequately protect the private data it held [101,102]. This breach is an extreme example of the costs, consequences and implications of inadequate data governance in a world increasingly reliant on the collection and use of data to develop AI. Stronger data governance policies (including both technical fixes like segmenting networks to isolate potential hackers and implementing robust data encryption, as well as external legislation creating stronger repercussions for consumer data loss) can help to prevent these types of breaches in future [101].

3.3 Open data sources and re-identification

The Australian Government has developed initiatives to better share and use reliable data sources. For instance, in the 2015 Public Data Policy Statement, the Government committed to, "optimise the use and reuse of public data; to release non sensitive data as open by default; and to collaborate with the private and research sectors to extend the value of public data for the benefit of the Australian public." [103] This announcement was backed up by several subsequent initiatives, culminating in the recent publication of three key reforms [104]:

- A new Consumer Data Right, whereby consumers can safely share their data with trusted recipients (e.g. comparison websites) to compare products and negotiate better deals [105]. The Consumer Data Right is a right for consumers to consent to share their data with businesses – there is no 'implied consent' for data transfers following the initial sharing. Consumers will also be able to keep track of and revoke their consent [105].
- A National Data Commissioner will implement and oversee a simpler, more efficient data sharing and release framework. The National Data Commissioner will be the trusted overseer of the public data system.
- New legislative and governance arrangements will enable better use of data across the economy while ensuring appropriate safeguards are in place to protect sensitive information.

In addition to these reforms, tens of thousands of government datasets are available to the public through the data.gov.au website. This resource is one reason why Australia scores very highly on the international Open Data Index (which measures government transparency online) [106] and may also be useful in catalysing AI innovation and development using rich and diverse Australian datasets.

The publication of non-sensitive data is imperative to support research and innovation, but there are ethical issues to consider. Many of these forms of data could alone be considered de-identified or non-personal, but the ability of AI to detect patterns and infer information could mean that individuals are identified from non-personal data. This information can be exploited in unethical ways that infringe on the right to privacy.

3.3.1 Case study: Ensuring privacy of de-identified data

In 2016, a dataset that included de-identified health information was uploaded to data.gov.au. It was expected that the data would be a useful tool for medical research and policy development. Unfortunately, it was discovered that in combination with other publicly available information, researchers were able to personally identify individuals from the data source. Quick action was taken to remove the dataset from data.gov.au.

The use of AI enabled devices and networks that can collate and predict data patterns has heightened the risk of being able to identify individuals in what was considered a de-identified dataset.

A report from Australia's Privacy Commissioner outlined the issues involved in the de-identification process of the data release and proposed the use of rigorous risk management processes, with clear documentation of the decision processes guiding the open publication of de-identified data [20].

The Government's Privacy Amendment (Re-identification Offence) Bill 2016 seeks to respond to a gap identified in privacy legislation about the handling of de-identified personal information by making it an offence to deliberately re-identify publicly released, de-identified government information.

The De-Identification Decision Making Framework by the Office of the Australian Information Commissioner and CSIRO can also assist in making decisions about these datasets.

Continued vigilance is required to ensure that de-identified datasets, that can be so useful to researchers, are adequately protected.

3.3.2 Case study: Locating people via geo-profiling

A recently published paper uses geo-profiles generated by publicly available information to (possibly) identify the artist Banksy, who has chosen to remain anonymous [107]. The study was framed as an investigation of the use of geo-profiling to solve a "mystery of modern art." The authors suggest that these methods could be used by law enforcement to locate terrorist bases based on terrorist graffiti. However, the ability of AI techniques to take publicly available data and make very personal inferences about individuals poses a significant ethical issue about privacy and consent issues even when dealing with publicly available, de-identified and non-personal data.

Any evaluation of the identifiability of data should examine how non-personal data will be shared and with whom. It should also consider how non-personal data could be used in conjunction with other data about the same individual. The Office of the Australian Information Commissioner has published a best practice guide on the use of data analytics, which clearly outlines considerations and directives to ensure effective data governance in line with the Australian Privacy Act [108]. In particular, the guide promotes the use of privacy-by-design to ensure that privacy is proactively managed and addressed through organisational culture, practices, processes and systems.

3.4 Bias in data

Machine learning and various other branches of AI are reliant on rich and diverse data sources to effectively train algorithms to create an output. If the training data does not include a robust, inclusive sample, bias can creep in, resulting in AI outputs with implicit bias that can disadvantage or advantage certain groups. Biased data inputs can lead to discrimination, most often against already vulnerable minority populations. One of the most basic requirements for preventing bias is controlling the data inputs to ensure they are appropriate for the AI systems that they are used to train. Unbiased datasets, too, can yield unfair results. This is explored more in chapter 5.3.

But simply using any and all input data is not a solution either, as the case study below demonstrates.

3.4.1 Case study: The Microsoft chatbot

Tay the Twitter chatbot was developed by Microsoft as a way to better understand how AI interacts with human users online. Tay was programmed to learn to communicate through interactions with Twitter users – in particular, its target audience was young American adults. However, the experiment only lasted 24 hours before Tay was taken offline for publishing extreme and offensive sexist and racist tweets.

The ability for Tay to learn from active real-time conversations on Twitter opened the chatbot up to misuse, as its ability to filter out bigoted and offensive data was not adequately developed. As a result, Tay processed, learned from and created responses reflective of the abusive content it encountered, supporting the adage, 'garbage in, garbage out' [109].

In addition to controlling the data inputs, consideration must be given to the impact of indirect discrimination. Indirect discrimination occurs as a result of the use of data variables highly correlated with other variables that can lead to discrimination [110]. For example, the neighbourhood that an individual lives in is often highly correlated with their racial background, and the use of this data to make decisions can thereby lead to racial bias.

3.4.2 Case study: Amazon same-day delivery

Amazon recently rolled out same-day delivery across a select group of American cities. However, this service was only extended to neighbourhoods with a high number of current Amazon users. As a result, predominantly non-white neighbourhoods were largely excluded from the service. The disadvantage to the neighbourhoods excluded from same-day delivery further marginalised communities that are likely to already be facing the impact of bias and discrimination.

Amazon could convincingly argue that it made its decision about where to roll out same-day delivery based on logistical and financial requirements. It did not intend to exclude non-white minorities, but because racial demographics tend to correlate with location, the decision did result in indirect discrimination.

The above example demonstrates the need for critical assessment of bias in data inputs used to make decisions and create outputs (whether for AI or otherwise). In scientific research and AI programming, strategies have been developed to optimise data inputs and sampling and reduce the impact of bias [110]. These issues need to be addressed at the development stage of AI, and as such, developers need to be able to appropriately assess their data inputs. Training and education systems that support the skills required to address bias in sampling data would help to address this ethical issue.

3.5 Key points

- Data governance is crucial to ethical AI; organisations developing AI technologies need to ensure they have strong data governance foundations or their AI applications risk being fed with inappropriate data and breaching privacy and/or discrimination laws.
- Organisations need to carefully consider meaningful consent when considering the input data that will feed their AI systems.
- The nature of the input data affects the output. Indiscriminate input data can lead to negative outcomes, this is just one reason why testing is important.
- AI offers new capabilities, but these new capabilities also have the potential to breach privacy regulations in new ways. If an AI can identify anonymised data, for example, this has repercussions for what data organisations can safely use.
- Organisations should constantly build on their existing data governance regimes by considering new AI-enabled capabilities and ensuring their data governance system remains relevant.

4 Automated decisions

> “Big Data processes codify the past. They do not invent the future. Doing that requires moral imagination, and that’s something only humans can provide. We have to explicitly embed better values into our algorithms, creating Big Data models that follow our ethical lead. Sometimes that will mean putting fairness ahead of profit.”
>
> Weapons of Math Destruction, Cathy O’Neil

Humans are faced with tens of thousands of decisions each day. These decisions can be influenced by our emotional state, fatigue, interest in the topic, internal biases and external influences. The decisions we make in a professional setting have the potential to significantly affect the greater community – for example, an insurance adjustor’s decision about a claim, a judge’s decision about a legal case or a banker’s decision about a loan application can have life-changing consequences for individuals.

Nationally and globally, AI is being used to guide decisions in government, banking and finance, insurance, the legal system and mining sectors.

The number of decisions driven by AI will likely grow dramatically with the development and uptake of new technology. When used appropriately, automated decisions can protect privacy, reduce bias, improve replicability and expedite bureaucratic processes. Australia’s challenge lies in developing a framework and accompanying resources to aid responsible development and use of automated decision technologies.

4.1 Humans in the loop (HITL)

Automated decisions require data inputs which are analysed and assessed against criteria to create data outputs and make decisions (Figure 5). During the design process each of these steps requires evaluation and assessment to ensure that the system performs as intended.

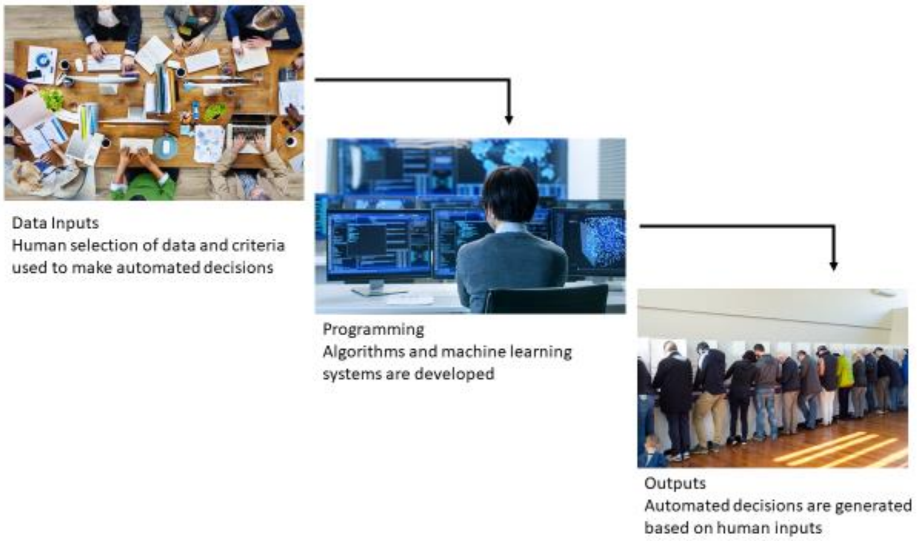

Data Inputs
Human selection of data and criteria used to make automated decisions
Programming
Algorithms and machine learning systems are developed
Outputs
Automated decisions are generated based on human inputs

Figure 5. Infographic showing three phases in developing automated decision systems

The concept of 'humans in the loop' (HITL) was developed to ensure that humans maintain a supervisory role over automated technologies [111]. HITL aims to ensure human oversight of exception control, optimisation and maintenance of automated decision systems to ensure that errors are addressed and humans remain accountable.

Automated decisions that affect large and diverse groups of people require the design principle of 'society in the loop' (SITL) [111] . For example, as discussed in Chapter 6.1, the automation of cars and the decisions and predictions they make will have far reaching effects on society as a whole. Incorporating SITL when designing automated vehicle systems involves considering the behaviour expected from the technology, and how it aligns with the goals, norms and morals of various stakeholders.

Both HITL and SITL promote careful consideration of the programming used to generate automated decisions to ensure that they reflect our laws, adhere to our human rights and social norms, as well as protect our privacy. Problems can occur when AI is developed without critical assessment and monitoring of the inputs, algorithms and outputs. Article 22 of the EU's GDPR law states that human beings have the right to not be subject solely to automated decisions when the decisions have legal ramifications [64].

4.2 Black box issues and transparency

Various branches of artificial intelligence studies, such as "Explainable AI" and "Transparent AI", seek to apply new processes, technologies and even other layers of AI to existing AI programs in order to make them understandable to users and other programmers [112].

This emerging area of research is likely to provide useful tools in understanding automated decision-making mechanisms. However, it is important to consider that transparent systems can still operate with a high error rate and significant bias. In addition, attempts to explain all of an algorithm's processes can slow down or hamper its effectiveness, and the explanations may still be far too complex. As an analogy, if asked why you chose a particular food, it wouldn't be helpful to explain the chemical processes that occurred in your brain while making that decision.

The question then becomes: what do we need to know about this algorithm to keep it accountable and functioning according to laws, rights and social norms? Any effective approach to regulating AI will need to be based on appropriate levels of transparency and accountability, which act in service of broad principles-based objectives.

4.2.1 Case study: Houston teachers

A proprietary AI system was used by the Houston school district to assess the performance of their teaching staff. The system used student test scores over time to assess the teachers' impact. The results were then used to dismiss teachers deemed ineffective by the system [113].

The teacher's union challenged the use of the AI system in court. As the algorithms used to assess the teacher's performance were considered proprietary information by the owners of the software, they could not be scrutinised by humans. This inscrutability was deemed a potential violation of the teachers' civil rights, and the case was settled with the school district withdrawing the use of the system. Judge Stephen Smith stated that the outputs of the AI systems could not be relied upon without further scrutiny, as they may be "erroneously calculated for any number of reasons, ranging from data-entry mistakes to glitches in the computer code itself. Algorithms are human creations, and subject to error like any other human endeavour" [113] .

With the increasing development and uptake of decision support systems and automated decision systems, similar cases will likely emerge in Australia in future – as such, it is important to understand the ethical issues inherent in automated decision-making, and consider methods of addressing these. Resolving issues of transparency associated with complex deep learning AI systems and proprietary systems will require multi-disciplinary input. Recourse mechanisms are one viable option to protect the interests of people affected by the use of automated decision systems – for instance, some countries have implemented the right to request human review of automated decisions. In the UK, when a decision is generated solely by automated processing, the subject of the decision must be made aware of this and has the ability (within a month of being notified) to lodge a request to "(*i) reconsider the decision, or (ii) take a new decision that is not based solely on automated processing as a recourse when autonomous decisions are used*" [93].

4.3 Automation bias and the need for active human oversight

Automation bias is "the tendency to disregard or not search for contradictory information in light of a computer-generated solution that is accepted as correct" [114] . Relying on automated decisions in situations where they cannot provide a consistently reliable outcome can result in increased errors by commission or omission (following, or failing to, act on advice from an automated decision system in error, respectively) [115]. These issues are particularly important when using automated decisions in situations requiring discretion.

Good decision making based on advice from automated decision systems requires the humans involved to exercise active thinking and involvement, rather than passively allowing automated decision systems to handle all of a task they are not suited for. When operators grow too reliant on automated systems and cease to question the advice they are receiving, problems can emerge, as demonstrated in the Enbridge pipeline leak of 2010.

4.3.1 Case study: automation bias and the Enbridge pipeline leak

In July 2010, a pipeline carrying crude oil ruptured near the Kalamazoo River in the US state of Michigan. The resulting clean-up took over five years at a cost of over US $737 million [116]. The disaster prompted numerous inquiries as well as academic papers. The large amount of environmental damage was caused by the delayed reaction to the rupture, which allowed oil to pump into the surrounding area for over 17 hours. During that time, there were two more "startup" moves to pump more oil, with the entire incident releasing 843,444 gallons of oil into the area.

An automated system did provide warnings to control centre personnel. A review into the incident found that operators had heard the alarms from the system and seen the abnormally low amounts of oil reaching the destination, but they had incorrectly attributed these warning signs to that planned shutdown. Reviews of the incident found that it was not until an outside caller notified them of the leak that it was discovered and action was taken [116].

Academics in 2017 analysed the disaster and suggested that regulators had overlooked complacency as a key driving factor. They suggested that "Industry, policy makers, and regulators need to consider automation bias when developing systems to reduce the likelihood of complacency errors [117]."

While the recommendations from review bodies highlighted the poor management of the incident, the academics pointed out that the people involved in the incident were all very experienced. They detailed earlier incidents in which more senior staff were more likely to overlook dangerous safety risks than junior ones [117]. They also pointed out that there had been frequent alarms in the past due to column separation problems, which were resolved by pumping more oil down the line to clear the track.

Thus in this case, experienced staff recommended a course of action that made the problem worse. This is a difficult problem to resolve, because as the academics point out: "According to some researchers, a problem such as occurred in the Enbridge case cannot be addressed by better training, given that it is human nature to ignore frequent false alarms [117]."

Designers of automated systems need to carefully consider the way their systems will interact with human operators, in order to counteract the damaging effects of overreliance and complacency, and ensure that the recommendations provided by the system cannot easily be misinterpreted in harmful ways. This will mean careful consideration of ways to ensure HITL design principles are implemented.

4.4 Who is responsible for automated decisions?

As AI systems are further developed and more widely applied, it will be important to create policy outlining where responsibility falls if things do go wrong. As an AI system has no moral authority, it cannot be held accountable in a judicial sense for its decisions and judgements. As such, a human must be accountable for the consequences of decisions made by the AI.

The main question arising from liability decisions is: which entity behind the technology is ultimately responsible, and at which point can a line be drawn between them? A recent Cambridge public law conference highlighted the complexity of who is responsible when something goes wrong with automated administrative decisions [118]:

- "To whom has authority been delegated, if that is indeed the correct analysis?
- Is it the programmer, the policy maker, the authorised decision-maker, or the computer itself?
- Is the concept of delegation appropriately used in this context at all? After all, unlike human delegates, a computer programme can never truly be said to act independently of its programmer or the relevant government agency?
- What if a computer process determines some, but not all, of the elements of the administrative decision? Should the determination of those elements be treated as the subject of separate decisions from those elements determined by the human decision-maker?"

4.4.1 Case study: Automated vehicles

In 2018, an Arizona pedestrian was killed by an automated vehicle owned by Uber. A preliminary report released by the National Transportation Safety Board (NTSB) in response to the incident states that there was a human present in the automated vehicle, but the human was not in control of the vehicle when the collision occurred [119]. There are various reasons why the collision could have occurred, including poor visibility of the pedestrian, lack of oversight by the human driver, and inadequate safety systems of the automated vehicle.

The complexities of attributing liability in instances of collisions involving automated vehicles are well documented [120]. In this case, although the legal matter was settled out of court and details have not been released, the issue of liability is complex as the vehicle was operated by Uber, under the supervision of a human driver and operated autonomously using components designed by various other tech companies. Following their full investigative process, the NTSB will release a final report of the incident identifying the factors that contributed to the collision.

The attribution of responsibility in regards to AI poses a pressing dilemma. There is a need for consistent and universal guidelines, applicable across various industries utilising technology that is able to make decisions significantly affecting human lives. In addition, policy may provide a universal framework that aids in defining appropriate situations where automated decisions and judgements may be used.

4.5 Key points:

- Existing legislation suggests that for government departments, automated decisions are suitable when there is a large volume of decisions to be made, based on relatively uniform, uncontested criteria. When discretion and exceptions are required, automated decision systems are best used only as a tool to assist human decision makers—or not used at all. These requirements are not mandated for other organisations, but are a wise approach to consider.
- Consider human-in-the-loop (HITL) principles during the design phase of automated decisions systems, and ensure sufficient human resources are available to handle the likely amount of inquiries.
- There must be a clear chain of accountability for the decisions made by an automated system. Ask: Who is responsible for the decisions made by the system?

5 Predicting human behaviour

> "Criminal sentences must be based on the facts, the law, the actual crimes committed, the circumstances surrounding each individual case, and the defendant's history of criminal conduct. They should not be based on unchangeable factors that a person cannot control, or on the possibility of a future crime that has not taken place."
>
> Former US Attorney General, Eric Holder

In a similar manner to which AI systems are able to process information and use it to make decisions, they are also able to extrapolate information and recognise patterns that can be used to make predictions about future behaviours and events. Although the previously discussed concepts of HITL, transparency and black box issues and accountability apply, the capability to predict future potential actions poses additional specific ethical concerns related to bias and fairness that require consideration.

When used appropriately, AI-enabled predictions can be powerfully accurate, replicable and efficient. They can be used to our advantage in place of human generated judgements and predictions, which can be subject to various extraneous variables that can be thought of as noise, such as bias, fatigue and effort. This technology could be especially useful in industries that require decision makers to generate frequent, accurate and replicable predictions and judgements such as the areas of justice, policing and medicine.

To appropriately assess the ethical issues associated with the use of AI enabled predictive and judgement systems it is crucial to first acknowledge the inherent issues associated with human judgements and predictions.

The Australian legal system uses precedent and sentencing guidelines to regulate decision making in an effort to combine discretion and address the influence of bias. Although judges spend years training they may still be impacted by cognitive biases, personal opinions, fluctuations in interest, fatigue and hunger [121].

Policies should promote Australia's colloquial motto, everyone deserves a fair go. Ensuring that AI systems are operating in a fair and balanced ways across the diverse Australian population is a cornerstone of ethical AI. Establishing industry standards supported by up-to-date guidelines could provide a baseline level of assessment for all AI used in Australia to support the use of fair algorithms.

5.1.1 Case study: Israeli Judges and decision fatigue

In one high profile study from Israel, academics examined parole hearings in Israel to determine what factors were most likely to result in a favourable ruling [122,123]. After observing 1,112 rulings, the researchers found that early in the day, and right after food breaks, the judges were far more likely to grant parole. The difference was extreme, with over 65% of cases right after rest breaks receiving a favourable ruling, compared to 0% just before a break.

The researchers suggested that the reason for this was that the judges simply became hungry and tired, resulting in harsher sentences. Other researchers have disputed the findings, indicating that the effect in this particular study was more likely due to other factors such as prospective parolees only having legal counsel at some times of day, or cases being deferred [124].

The extent to which decision fatigue and general exhaustion affects judicial decisions is open to debate, but the problems associated with decision-fatigue are well supported in academic literature, as are the difficulties in grappling with cognitive biases. Miscarriages of justice are frequently attributable to human error or misconduct.

Even the best decision-makers will sometimes resort to mental shortcuts, or “heuristics” to understand a situation and make decisions [125]. Well-designed AI has the ability to perform in these situations with a much higher degree of replicability and consistency than humans.

5.2 Bias, predictions and discrimination

Indirect discrimination occurs when data variables that are highly correlated with discriminatory variables are included in a model [110] – to give an example, an algorithm might not explicitly consider race as a factor, but it might discriminate against a neighbourhood filled almost entirely with people of one race, leading to a similar result. The superior ability of AI to recognise patterns creates serious potential ethical issues when it is used to make predictions about human behaviour. To ensure that predictive systems are not indirectly biased, all variables used to develop and train the algorithms must be rigorously assessed and tested. In cases with higher risk, it may be important to run smaller tests or simulations before using them on the broader public. In addition, the model itself should be assessed and monitored to ensure that bias does not creep in.

Australian legislation prohibits discrimination (unfair or unequal treatment of a group or individual) on the basis of race, colour, sex, religion, political opinion, national extraction, social origin, age, medical record, criminal record, marital or relationship status, impairment, mental, intellectual or psychiatric disability, physical disability, nationality, sexual orientation, and trade union activity [126].

AI is set to use many of these indicators to make predictions about health or behaviour. Not all of these will constitute discrimination. Race, for example, may prove to be a relevant indicator of a particular health problem that could be avoided—consider fair skinned people, who are more at risk of skin cancer [127]. An AI that assesses skin cancer risk would need to take into account skin tone as a factor.

So when is the use of a particular input variable considered discrimination? Careful consideration will need to be given as to what kind of outcomes constitute discrimination. This will be helpful in considering what variables can be included, but even beyond explicit variables there are also indirect ones.

Researchers have pointed out that there are many indicators, such as postcodes, education and family history that can effectively indicate race without it explicitly needing to be included as an indicator. In one example from the US, geographic information was used to determine the cost of test preparation services. It was revealed that this method unfairly discriminated against Asian American students who were charged higher fees for academic thesis review services than other non-Asian students [128]. Although ethnicity was not specifically considered in the pricing structure of the service, the use of location based pricing disproportionally impacted Asian-American students with higher fees resulting in indirect discrimination.

This also prompts another ethical question for consideration: beyond racial discrimination, should location-based discrimination be permissible or is this still discrimination?

The issue of racial bias has been exposed in AI used in the US, in courts to assess the likelihood that someone will re-offend, and by police to focus on crime hotspots and identify potential suspects. Research and debate is already occurring on the suitability of these tools in the Australian context [129,130].

5.2.1 Case study: The COMPAS system and sentencing

The COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) system is currently used in many US courts to advise judges on sentencing and probation decisions. The system evaluates individuals using 137 questions and assigns a risk score out of ten that indicates the probability of how likely it is that a person will re-offend.

Although race is not directly assessed by the system, zip code is. Research by non-profit media outlet ProPublica found that black people profiled by COMPAS were twice as likely to be incorrectly labelled as being at high risk of committing violent repeat offences than white people [16].

The creators of the algorithm, Northpointe, responded to the report denying that their algorithm is biased against black people [17]. Northpointe indicated that across their risk scores, ranging from one to ten, there were equal levels of recidivism from black and white groups and that there was no racial bias in their system [23].

But the problem, according to ProPublica, was not in the correct predictions, which they acknowledged were reasonably fair across the two racial groups. The problem was in the incorrect predictions (or false positives).

“When we analysed the 40 percent of predictions that were incorrect, we found a significant racial disparity. Black defendants were twice as likely to be rated as higher risk but not re-offend. And white defendants were twice as likely to be charged with new crimes after being classed as lower risk.” [18]

Both ProPublica and Northpointe were able to examine the same system and data and come up with opposing findings on the racial bias of the system. The discrepancy between the two analyses essentially came down to the way each group assessed and measured fairness and balanced it with accuracy of the system. It also suggests that internal reviews may miss key problems if they base their analysis on the same assumptions of fairness as the original design.

The ability to assess bias in predictive systems is intrinsically linked with how fairness is measured along with the level of transparency involved.

The way that the COMPAS system weighs and assesses the defendant’s input variables to calculate their risk score is proprietary software, so assessment of the way that the system deals with racial indicators cannot be directly assessed. This lack of transparency presents significant ethical issues as the predictive scores assigned to individuals can have significant effects on their lives. This lack of clarity on how the system works has already resulted in one challenge being filed with the US Supreme Court on the basis that due to a lack of transparency, the accuracy and validity of COMPAS could not be disputed. The case was not heard [131].

The complex interactions between bias, fairness and transparency in AI enabled predictive systems are exemplified in the COMPAS case study. Although the use of AI enabled predictive systems to help support decision making processes poses great potential benefits in boosting replicability and reducing human error and bias, there are inherent ethical issues that must be addressed, particularly the effects of indirect discrimination and fairness. With appropriate transparency and accountability guidelines the use of these systems could allow us to examine the way the predictions are made in turn giving us the ability to make adjustments to address bias.

There may need to be ongoing conversations with the community about how their data is used. COMPAS, for example, may not explicitly consider race, but it does explicitly consider the family background of the offender and whether their parents are married or separated [132]. Would this be considered fair?

Australia must also contend with different sentencing outcomes for racial groups. In NSW a Suspect Target Management Program (STMP) was introduced in 2000 to identify and target repeat offenders to proactively disrupt their criminal behaviour. This program has come under some criticism for disproportionally targeting Aboriginal people.

Finding the right balance between proactively preventing crime and inaccurate group profiling is an ongoing ethical challenge across all jurisdictions.

The designers of algorithms need to pay careful attention to how their systems come to a prediction and there may be a role for government bodies in determining general boundaries and review and monitoring processes—based on existing laws regarding discrimination—for how issues relating to the separate but related issues of bias and fairness can best be addressed and mitigated.

5.3 Fairness and predictions

The challenge of ensuring fairness in algorithms is not limited to biased datasets. The input data comes from the world, and the data collected from the world is not necessarily fair.

The various population sub groups, such as men, women or people of a particular race or disability, are all likely to be represented differently across datasets. When put into an AI or automated system, various sub-populations will rarely if ever show exactly the same input and output results. This makes it likely that the AI will be inherently discriminatory in one way or another—and that is assuming that the input information is reasonably accurate.

Accuracy in datasets is rarely perfect and the varying levels of accuracy can in and of themselves produce unfair results—if an AI makes more mistake for one racial group, as has been observed in facial recognition systems [133], that can constitute racial discrimination.

Then, there is the issue of relatively accurate data that represents and unfair situation in the real world. Disadvantaged groups may be disadvantaged for historical or institutional reasons, but that information isn't necessarily understood or compensated for by an AI, which will assess the situation based on the data alone.

So what is "fairness?" It really depends who you ask.

"Fairness" is a difficult concept to pin down and AI designers essentially have to reduce it to statistics. Researchers have come up with many dozens of mathematical definitions to define what fairness means in an algorithm and many of them perform extremely well when measured from one angle, but from a different angle can produce very different results. This concept of differing perspectives of fairness is exemplified in the COMPAS case study. ProPublica and the Northpointe had diverging perspectives on how to accurately judge parity between the assessment of white and black defendants resulting in completely different answers as to whether the system was biased.

Put simply: it will sometimes be mathematically impossible to meet every single fairness measure because some of them contradict each other and multiple datasets will be used in systems, and these datasets will almost never be exactly equal in accuracy or representativeness. Tradeoffs will sometimes be necessary.

It is important for government and society to give consideration to the degree of flexibility that designers of AI systems should have when it comes to making trade-offs between fairness measures and other priorities like profit. There needs to be serious consideration given to whether the net benefits of the algorithm justify its existence, and whether it is justified in the ways it treats different groups.

Companies and consumers will be faced with decisions to make about which algorithms best represent their values. If a company prioritises profit ahead of various forms of fairness, can they justify it? A key consideration will be how transparent they are in this decision so the broader public are able to make informed choices, and companies are acting in accordance with public expectations.

5.3.1 Case study: The Amazon hiring tool

In 2014, global e-commerce company Amazon began work on an automated resume selection tool. The goal was to have a system which could scan through large numbers of resumes, and determine the best candidates for positions, expressed as a rating between one and five stars. The tool was given information on job applicants over the preceding ten years.

In 2015, the researchers realised that because of male dominance in the tech industry, the tool was now assigning higher ratings to men than women. This was not just a problem of there being larger numbers of qualified male applicants—the key word "women" appearing in resumes resulted in them being downgraded.

The tool was designed to look beyond the common keywords, such as particular programming languages, and focus on more subtle cues like the types of verbs used. Certain favoured verbs like "executed" were more likely to appear on the resumes of male applicants [134].

The technology not only unfairly advantaged male applicants, it also put forward unqualified applicants. The hiring tool was scrapped, likely due to these problems.

Talent websites such as LinkedIn use algorithms in their resume-matching systems, but key staff have said that AI in its current state should not be making final hiring decisions on its own, and instead should be used as a tool for human recruiters.

In Australia, ANZ bank in 2018 indicated it is researching the use of AI in hiring practices [135]. It states that the goal is to find candidates in a "fair and unbiased" way. Algorithms used in such a manner can take decisions out of human hands, but as the Amazon case demonstrates, this does not necessarily mean they are without human bias, especially in the training data. Hiring tools such as these will need to ensure that the data inputs—which in this case, includes measures like the time taken to answer certain questions—are relevant and reflective of actual performance in the role. Testing this kind of outcome is difficult, as metrics for employee performance are not always easily expressed in KPIs. There may indeed be potential for ethical outcomes from these tools, but it should not be assumed that these tools will be more ethical or less biased than human recruiters.

5.4 Transparency, policing and predictions

Predictive analytics powered by big data can boost the accuracy and efficiency of policing in Australia, but lessons learned overseas point to a need for strong transparency measures and for the public to be actively involved in any program. Predictive policing programs are active in the US and in the UK, though there is little peer-reviewed material on their effectiveness [24].

Most predictive policing tools are not about predicting a specific crime—instead they can be used to either profile people or places, and measure the effectiveness of particular policing initiatives. They aim to inform police on crime trends and what is or isn't working, and focus on long-term persistent problems rather than individual crimes [24].

5.4.1 Case study: Predictive policing in the US

When data analysis company Palantir partnered up with the New Orleans police department, it began to assemble a database of local individuals to use as a basis for predictive policing. To do this it used information gathered from social media profiles, known associates, licence plates, phone numbers, nicknames, weapons and addresses [136]. Media reports indicated that this database covered around 1% of the population of New Orleans. After it had been in operation for six years there was a flurry of media attention over the “secretive” program. The New Orleans Police Department (NOPD) clarified that the program was not secret and had been discussed at technology conferences. But insufficient publicity meant that even some city council members were unaware of the program.

A “gang member scorecard” became a focus of media coverage, and groups such as the American Civil Liberties Union (ACLU) pointed out that operations like this require more community approval and transparency. The ACLU argued that the data used to fill out these databases could be based on biased practices, such as stop-and-frisk policies which disproportionately target African American males, and this would feed into predictive policing databases. [136] This could mean more African Americans were scanned by the system, creating a feedback loop in which they were more likely to be targeted. This should be a key consideration in any long-term use of an AI program—is the data being collected over time still serving the intended outcome, and what measures ensure that this is being regularly assessed?

The NOPD terminated its agreement with Palantir and some defendants that were identified through the system have raised the use of this technology in their court defence and attempted to subpoena documents relating to the Palantir algorithms from the authorities [137].

The Los Angeles Police Department (LAPD) still has an agreement with Palantir in relation to its LASER predictive policing system. The LAPD also runs Predpol, a system which predicts crime by area, and suggests locations for police to patrol based on previously reported crimes, with the goal of creating a deterrent effect. These programs have both prompted pushback from local civic organisations, who say that residents are being unfairly spied upon by police because their neighbourhoods have been profiled [138], potentially creating another feedback loop in which people are more likely to be stopped by police because they live in a certain neighbourhood, and once they have been stopped by police they are more likely to be stopped again.

In response, local police have invited reporters to see their predictive policing in action, arguing that it also helps communities affected by crime and pointing out that early intervention can save lives and foster positive links between police and entire communities [138]. Some police officers were also cited in media pointing out that there needs to be sufficient public involvement and understanding and acceptance of predictive policing programs in order for them to effectively build those community links.

There are clear ethical issues that arise with the advent of predictive policing. One of the key issues is the need for transparency in how these systems work so that they can be adequately assessed and that they remain accountable to the citizens affected by them.

In addition the exploitation of potentially personal data by these systems could infringe on privacy rights accorded to Australians through the Privacy Act. The trade-off between the increased ability of the police to prevent and monitor crime and the protection of personal privacy is discussed in Chapter 6.2.

5.4.2 Case study: Predictive policing in Brisbane

One predictive policing tool has already been modelled to predict crime hotspots in Brisbane [139]. Using 10 years of accumulated crime data, the system used 70% of the data to predict crime, with the researchers seeing if its predictions correlated with the remaining 30%. The results proved more accurate than existing models, with an improvement of 16% accuracy for assaults, 6% more accuracy for predicting unlawful entry, 4% better accuracy for predicting drug offences and theft, and 2% better for fraud [139,140]. The system can predict long term crime trends, but not short term ones [140]. The Brisbane study used information from location-based app foursquare, and incorporated information from both Brisbane and New York.

Predictive policing tools typically use four broad types of information [140]:

- Historical data, such as the long term crime patterns recorded by police as crime hotspots.

- Geographical and demographic information, including distances from roads, median values of houses, marriages, socioeconomic and racial makeup of an area.
- Social media information, such as tweets about a particular location and keywords
- Human mobility information, such as mobile phone usage and check ins and the associated distribution in population

The Brisbane study primarily used human mobility information. Further research is needed, but this study suggests that some types of input information may be more effective at gauging accurate information than others—so careful consideration may warrant emphasizing some types of input information over others, given that they will have varying impacts on privacy and some less intrusive approaches may be less likely to provoke public distrust. Public trust, as case studies from the US demonstrate, is a crucial element of the effectiveness of predictive policing tools.

The debate over the use of AI in policing is ongoing. One clear outcome has been that if new technologies are used in law enforcement without public endorsement, then those systems will not effectively serve the police or general public. Transparency about how these systems operate and in what circumstances they will be used are at the core of ensuring they remain ethical and accountable to the communities they protect. It may be constructive for agencies to consider public engagement strategies and feedback mechanisms before introducing new AI technologies that will significantly affect the public. It may also be prudent to consider risk analysis to provide objective information for the public on the beneficial outcomes of using AI enabled systems versus traditional policing methods.

5.5 Medical predictions

AI predictions have the ability to add immense value to the Australian health care system in patient management, diagnostics and care. However, like all new medical advances and methods, AI systems used in health care require close management and gold standard research before implementation.

5.5.1 Case study: Predicting coma outcomes

A program in China that analyses the brain activity of coma patients was able to successfully predict seven cases in which the patients went on to recover, despite doctor assessments giving them a much lower chance of recovery [141]. The AI took examples of previous scans, and was able to detect subtle brain activity and patterns to determine which patients were likely to recover and which were not. One patient was given a score of just seven out of 23 by human doctors, which indicated a low probability of recovery, but the AI gave him over 20 points. He subsequently recovered. His life may have been saved by the AI.

If this AI lives up to its potential, then this kind of tool would be of immense value in saving human lives by spotting previously hidden potential for recovery in coma patients—those given high scores can be kept on life support long enough to recover. But it prompts the question: what about people with low scores?

Rigorous peer-reviewed research should be conducted before such systems are relied upon to inform clinical decisions. Ongoing monitoring, auditing and research are also required.

Assuming the AI is accurate—and a number of patients with low scores would need to be kept on life support to confirm the accuracy of the system—then the core questions will revolve around resourcing. Families of patients may wish to try for recovery even if the odds of success are very small. It is crucial that decisions in such cases are made among all stakeholders and don’t hinge solely on the results from a machine. If resources permit, then families should have that option.

A sad reality of the hospital system is that every day, resources determine life and death decisions and those hospital resources are not limitless. If an AI system has the potential to direct those resources into situations with the best chance of recovery, then that is a net gain in lives saved.

In many ways, an AI-enabled diagnostic tool is no different from other diagnostic tools—an abnormal reading from an electrocardiography device has essentially made a similar prediction with life or death consequences, and assisted doctors in giving the best treatment they can to the highest number of people.

5.5.2 AI and health insurance

If artificial intelligence can deliver more accurate predictions in areas like healthcare, then this has ramifications for insurance. If an AI can assess someone's health more accurately than a human physician, then this is an excellent result for people in good health—they can receive lower premiums, benefiting both them and the insurance company due to the lower risk.

But what happens when the AI locates a hard-to-spot health problem and the insurer increases the premium or denies that person coverage altogether, in order to be able to deliver lower premiums to other customers or increase profit? In Australia, there are prohibitions on discriminating against people with pre-existing conditions, but it is permissible for insurers to impose a 12-month waiting period for any payouts for people with pre-existing conditions, to ensure they do not take out insurance just ahead of an expensive procedure [142]. Rules such as this one may become even more important moving forward.

Numerous medical technologies improve the ability to diagnose health problems and thus improve the ability to calculate risk, with genetic screening being just one recent example. AI technologies may boost the accuracy of these risk assessments, but they do not necessarily change the nature of insurance beyond this context.

In the event of a dramatic leap in the accuracy of health predictions, regulatory responses may be needed to ensure people with health problems are not priced out of the insurance market. At the present stage, absent a far-reaching shift in the provision of healthcare, these responses are likely to fit comfortably within existing legislation regarding the insurance industry.

5.6 Predictions and consumer behaviour

Whether they are aware of it or not, Australians who use social media or search engines are likely to have received targeted advertisements. These include adverts from platforms giants such as Google or Facebook that pitched a product based on their online activity. This can provoke mixed feelings among consumers.

Writing in a journal on consumer preferences, several scholars recently examined the ability of AI-enabled technologies to accurately predict consumer behaviour and provide targeted advertising. They state that what may be a short-term boon to advertisers can come at a cost beyond the immediate monetary input:

"We contend that some of the welfare-enhancing benefits of those technologies can backfire and generate consumer reactance if they undermine the sense of autonomy that consumers seek in their decision-making. That may occur when consumers feel deprived of their ability to control their own choices: predictive algorithms are getting better and better at anticipating consumers' preferences, and decision-making aids are often too opaque for consumers to understand" [143] .

There is a lot of misunderstanding on the part of consumers regarding the techniques used to determine their preferences for targeted advertising. When questioned by the US senate, Facebook CEO Mark Zuckerberg had to repeatedly deny that Facebook messenger listens to audio messages between people to better sell them adverts [144]. While Facebook does not appear to be mining audio messages, the company has utilised machine learning to predict when users might change brand as part of a "loyalty prediction" program offered to advertisers, and these claimed capabilities were only reported in media through leaked documents [145].

Advertising and data collection standards that address AI capabilities and ensure privacy is protected will be crucial in building trust between consumers and companies, while ensuring a balance between intrusive and beneficial targeted advertising.

5.6.1 Case study: Manipulating the mood states and perceptions of users

Controversial research, published in a peer-reviewed journal, used Facebook's platform to demonstrate that users' moods can be manipulated by filtering their feeds (comments, videos, pictures and web links posted by their Facebook friends). [146] Reduced exposure to feeds with positive content led to the user posting fewer positive posts, and the same pattern occurred for negative content. This study was publicly criticised for failing to gain informed consent from Facebook users. However, setting aside the issue of informed consent, the study highlights the power of AI-driven 'filtering practices' to shape user mood, which can be used to enhance the impact of targeted advertising.

AI technology can go beyond filtering news feeds, to manipulating video content in real time. A research project called Deep Video Portraits was recently showcased at an international conference on innovations in animation and computer graphics, and showed videos of 'talking heads' being seamlessly altered [147]. The technology produced subtle changes in emotion and tone that are difficult to distinguish from real footage. While the researchers contend that the technology can be used by the film industry, the technology also has implications for the fake news phenomenon. As one of the Deep Video Portraits researchers Michael Zollhofer [148] stated in a press release, "With ever-improving video editing technology, we must also start being more critical about the video content we consume every day, especially if there is no proof of origin."

These examples show that AI can be used to slant information without user knowledge, and for the purpose of influencing how consumers of media feel and perceive reality. AI-based manipulations of this calibre will require new controls to ensure users can trust the content they receive and are informed of advertising tactics.

One potential approach here could involve requirements for information to be posted on websites like Facebook revealing when AI techniques have been used to enhance targeted advertisements. This measure would be similar to the EU Cookie Law, which requires websites to explain to users what information is captured by the site and how it is used [149]. More sophisticated techniques might be required for fake videos. For example, the US Defence Advanced Research Projects Agency runs a program called Media Forensics that is developing AI tools to detect doctored video clips [150].

5.7 Key points

- AI is not driven by human bias but it is programmed by humans. It can be susceptible to the biases of its programmers, or can end up making flawed judgments based on flawed information. Even when the information is not flawed, if the priorities of the system are not aligned with expectations of fairness, then the system can deliver negative outcomes.
- Justice means that like situations should deliver like outcomes, but different situations can deliver different outcomes. This means that developers need to pay special care to vulnerable, disadvantaged or protected groups when programming AI.
- Full transparency is sometimes impossible, or undesirable (consider privacy breaches). But there are always ways to achieve a certain degree of transparency. Take neural nets, for example: they are too complex to open up and explain, and very few people would have the expertise to understand anyway. However, the input data can be explained, the outcomes from the system can be monitored, and the impacts of the system

can be reviewed internally or externally. Consider the system, and design a suitable framework for keeping it transparent and accountable. This is necessary for ensuring the system is operating fairly, in line with Australian norms and values.

- Public trust is of key importance. Organisations would be well advised to go beyond the “letter of the law” and instead follow best practice when designing AI.
- Not all input data is equally effective, but there are also varying levels of invasiveness. Carefully consider whether there are alternative forms of less invasive input data that could yield equal or better results. Ask: “Is there less sensitive data that could deliver the necessary results?”
- Know the trade-offs in the system you are using. Make active choices about them that could be justified in the court of public opinion. If a poorly-designed AI system causes harm to the public, ignorance is unlikely to be an acceptable defence.

6 Current examples of AI in practice

In addition to addressing the ethical issues that have arisen from data, automated decisions and predictive technologies it is important to consider examples of how AI systems are integrated in ways that have already or could potentially have an enormous impact on society. In this chapter we discuss the ethical issues associated with AI enabled vehicles and surveillance. These technologies have been a key area of focus in ethical AI discussions and are often used as examples of areas that need focussed attention from governments to help regulate their use.

6.1 Autonomous vehicles

> "There is a naïve view that AVs are in themselves beneficial. They can be beneficial only if we deliberately make them so"
>
> Fighting Traffic, Peter D. Norton

Autonomous vehicles (AVs) represent a major possibility for artificial intelligence applications in transport. However, the definition of 'autonomy' exists on a spectrum, rather than as a binary. There are five levels of vehicle autonomy, defined by the Society of Automotive Engineers [151] (Figure 6).

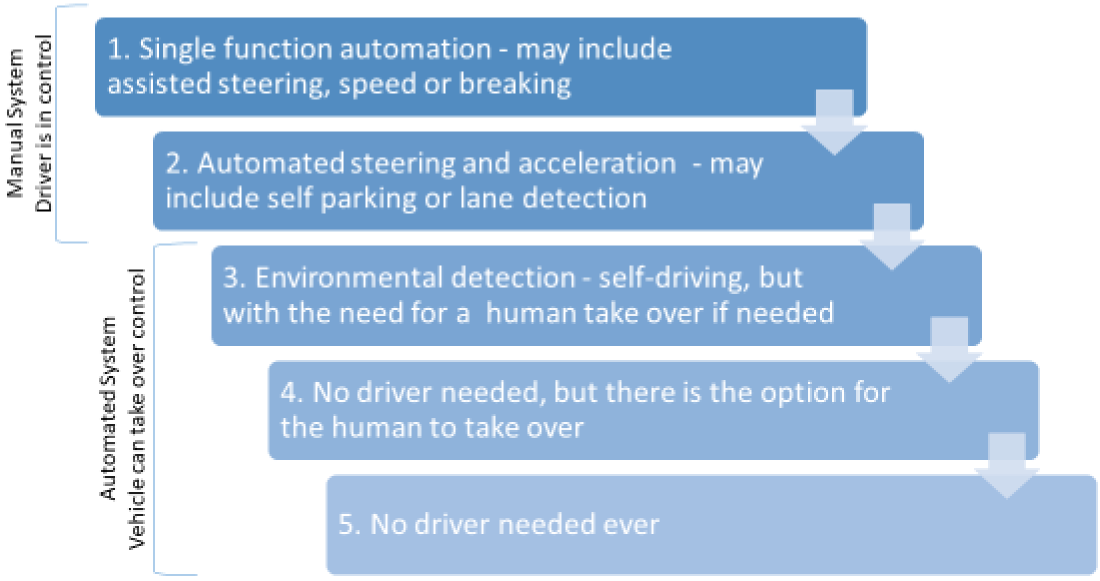

Manual System
Driver is in control
Automated System
Vehicle can take over control
1. Single function automation - may include assisted steering, speed or breaking
2. Automated steering and acceleration - may include self parking or lane detection
3. Environmental detection - self-driving, but with the need for a human take over if needed
4. No driver needed, but there is the option for the human to take over
5. No driver needed ever

Figure 6. Infographic showing the five levels of vehicle autonomy

Data source: Adapted from the Society of Automated Engineers [151]

6.1.1 Autonomous vehicles in Australia

In Australia, transport ministers have agreed to a phased reform program delivered by the National Transport Commission (NTC) to support the safe and legal operation of automated vehicles from 2020 [152].

The NTC recently published guidelines for trialling automated vehicles in Australia [153]. The guidelines describe an application process to request the trial of automated vehicles on Australian roads. The criteria that must be addressed in the applications include details of, the trial location, the technology being trialled, the traffic management plan, the infrastructure requirements of the trial, how the organisations will with engage with the public and how they will manage changes over the course of the trial. The guideline intends to support innovation, create a national set of guidelines and encourage investment in Australia as a *'global testbed for automated vehicles'* [153]. Level 4 automated vehicles trials are currently being run in Melbourne, Perth and Sydney.

The NTC also released a policy paper that helps clarify how traffic laws are applied to vehicles with automated functions at this point in time. In particular the paper clarifies the responsibility of the human driver *"for compliance with road traffic laws when a vehicle has conditional automation engaged at a point in time" [154]* . The most recent release by the NTC addresses the laws that need to be changed to support the use of automated vehicles [155].

The NTC is currently working on several more reports including the need to address insurance issues and safety and the regulation of vehicle data. In preparation for the changes automated cars will bring, the Minister for Infrastructure, Transport and Regional Development announced the upcoming opening of a new Office of Future Transport Technologies in October 2018 [156]. These upcoming initiatives will help provide the required foresight and support that will enable Australia to keep pace with the rapidly changing capabilities of automated vehicles.

6.1.2 Costs and benefits of automated vehicles

There is growing commercial interest around AVs, with sales predicted to reach 1 million by 2027 and 10 million by 2032 [157-161]. However, many artificial intelligence experts caution that Level 5 autonomy is still much further away than generally believed. As well as technological barriers, the affordability, capability and accessibility of AVs, along with privacy concerns, could also impact future uptake [162].

However, AVs – even without reaching Level 5 autonomy – also have the potential to deliver numerous social, environmental, and financial benefits. Research has found that AVs could ease congested traffic flow and reduce fuel consumption [163]; reduce travel costs (through lowering the cost of crashes, travel time, fuel, and parking) [164] ; and enable a smaller car fleet [165]. (However, the additional convenience and mobility afforded by AVs could also translate into greater demand for private vehicle travel over public transport, walking, or cycling [166-168]).

Safety represents another major potential benefit of vehicle automation. US research has found that over 90% of car crashes result from human error [169] and 40% of fatal crashes are caused by distraction, intoxication or fatigue [164]. Removing the human driver from the equation will therefore eliminate these incidents – some estimates suggest that full vehicle automation could reduce traffic accidents by up to 90% [170]. However, there are also safety concerns surrounding AVs, especially since the high-profile incident in March 2018 when an AV hit and killed a pedestrian [171]. (The preliminary report on the accident does not determine probable cause or assign culpability, but did note a number of design decisions that could be characterised as questionable, such as the fact that the vehicle operator monitors the self-driving interface via a screen in the car, but is also expected to apply emergency braking if necessary, and will not be alerted by the system if emergency braking is needed [119]). AVs also introduce the issue of cybersecurity, with the 2015 hacking of a Jeep Cherokee demonstrating the vulnerability of digitally connected cars [172].

6.1.3 Ethical principles and automated vehicles

AVs, as machines which have to make decisions (in accordance with programming determined by humans) are also subject to complex and difficult ethical considerations. Some key ethical questions surrounding AVs include:

- Should the car be programmed to take an egalitarian approach (maximising benefit for the highest number of people) or a negative approach (maximising benefit for the occupant only, and increasing risk for everyone else)? [173]
- Should car owners have a say in setting the car's 'moral code'? [173]
- In situations where harm to humans is unavoidable, would it be acceptable for AVs to perform some kind of prioritisation – e.g. based on age?
- How should AVs distribute the risk of driving – for instance, would it be acceptable to program a car that valued occupants over pedestrians, or vice versa?
- In instances such as the fatal AV crash of March 2018, who is responsible for the harm caused – the operator or the car manufacturer?

It is relatively straightforward to program AVs in accordance with certain rules (e.g. do not cross a lane boundary; do not collide with pedestrians, etc. – although, as the March 2018 crash shows, the technology is still far from perfect in following these rules). 'Dilemma situations' represent cases where not all rules can be followed, and some kind of decision has to be made in accordance with ethical principles. Usually, a 'hierarchy of constraints' is needed to determine action. This has prompted debate over how an autonomous vehicle should weigh outcomes that will cost human lives in various situations where an accident is inevitable.

Utilitarianism – maximising benefits and reducing harm for the greatest number of people, without distinction between them – is a strong principle underlying considerations of the ethics of AVs. Research by MIT has found that most people favour a utilitarian approach to AVs [174]. However, while participants approved of utilitarian AVs in theory and would like others to buy them, they themselves would prefer to ride in AVs that protect occupants at all costs [174]. Given that car manufacturers will therefore be incentivised to produce cars programmed to prioritise occupant safety, any realisation of utilitarian ethics in AVs will likely be brought about only through regulation [175].

Utilitarian principles are also complex to implement, and give rise to ethical conundrums of their own. For instance, following the principle of harm reduction, should an AV be programmed to hit a motorcyclist with a helmet instead of one without a helmet, since the chance of survival is greater? [176] Alternatively, it could be argued that AVs should, where possible, 'choose' to hit cars with greater crashworthiness – a development which would disincentivise the purchase of safer cars [177]. A 'consequentialist approach' that uses a single cost function (e.g. human harm) and encodes ethics purely around the principle of reducing that cost is therefore not broadly feasible [177].

Utilitarianism is not the only consideration in the ethics of AVs. A recent study surveyed millions of people across hundreds of countries to gauge moral preferences in AVs and what priorities they should have in the event of an unavoidable accident [30]. The researchers used an online survey to get over 39 million responses to hypothetical ethical dilemmas for AVs. The strongest preferences were for sparing human lives over animal lives, sparing more lives, and sparing young lives. The results indicated a popular preference for sparing the lives of children over adults. Notably, not all parts of the world saw eye-to-eye on how AVs should make such life-and-death decisions. In Eastern cultures, young lives and fit people were not given the same preference for protection as in Western cultures, while pedestrians were given extra weight [30]. Southern cultures expressed a stronger preference for protecting women [30].

6.1.4 Germany's Ethics Commission report

In 2017, Germany became the first country in the world to attempt to answer and codify some of these ethical questions in a set of formal ethical guidelines for AVs, drawn up by a government-appointed committee comprised of legal, technical, and ethical experts. The full report contains 20 propositions. Key among these are [1]:

- Automated and connected driving is an ethical imperative if the systems cause fewer accidents than human drivers (that is, a positive balance of risk).
- In hazardous situations, the protection of human life must always have top priority. The system must be programmed to accept damage to animals or property in a conflict if this means that personal injury can be prevented.
- In the event of unavoidable accident situations, any distinction between individuals based on personal features (age, gender, physical or mental constitution) is impermissible.
- In every driving situation, it must be clearly regulated and apparent who is responsible for the driving task: the human or the computer.
- Drivers must always be able to decide themselves whether their vehicle data are to be forwarded and used (data sovereignty).
- Genuine dilemma situations (e.g. the decision between human lives) depend on the actual specific situation and cannot be standardised or programmed. It would be desirable for an independent public sector agency to systematically process the lessons learned from these situations.
- In the case of automated and connected driving systems, the accountability that was previously the sole preserve of the individual shifts from the motorist to the manufacturers and operators of the technological systems and to the bodies responsible for taking infrastructure, policy and legal decisions.

One priority to keep in mind is the need for a uniform set of regulations for autonomous vehicles operating on Australian roads, which takes into account that Australian vehicles may be operating under different road rules than in the location manufactured—this has ramifications for which side of the road the vehicle drives on, or the presence of roundabouts. There is also the global context to consider—most road rules have a global context. This necessitates international collaboration.

The Australian Government has been an active participant in work in the UN World Forum for the Harmonization of Vehicle Regulations. These safety regulations are then incorporated into national law across many countries. In Australia they are known as Australian Design Rules under the Motor Vehicle Standards Act 1989.

In addition, another relevant UN working group is the Global Forum for Road Traffic Safety. The outcomes from this working group affect the rules made and implemented by Australian state governments. The Australian government is becoming more involved with this group as rules are being considered relating to autonomous vehicles.

Australia's vehicle safety regulations are already based on international standards, so in the longer term context of autonomous vehicles it is likely Australia will take up the appropriate standardised international safety frameworks. There would be some localised exceptions, such as local legislation on child seatbelts or rules relating to the supply of vehicles for the appropriate side of the road.

In the interim period, regulatory processes are being developed within Australia [178].

The NTC is currently in the process of considering automated vehicle liability issues [179]. It is also considering as regulation around safety assurance systems for AVs, with four potential reform options ranging from the baseline option of using existing regulation to manage safety, right through to the introduction of a regulatory system that is nationally managed from point of supply and while in service. This would impose a primary safety duty on the manufacturer of the automated driving system and require them to certify that their AVs adhere to safety criteria [180]. Interestingly, the report does not address ethical considerations in regards to AVs, stating that "safety dilemmas with ethical implications are already largely captured by the safety criteria". The safety criteria state that AVs must be able to:

- "detect and appropriately respond to a variety of foreseeable and unusual conditions affecting its safe operation, and to interact in a predictable and safe way with other road users (road users include other automated and non-automated vehicles and vulnerable road users)
- take steps towards achieving a minimal risk condition when it cannot operate safely
- prioritise safety over strict compliance with road traffic laws where necessary" [180]

Uniformity is also necessary because regardless of the specific priorities that are chosen in certain "life-or-death" scenarios, vehicles that are all using similar operating principles can more easily determine the safest way to respond in a given scenario. If different manufacturers are all creating autonomous vehicles that operate to different specifications, it is likely that safety would suffer.

6.2 Personal identification and surveillance

The ability of AI-enabled face, voice and even gait [181] recognition systems provide immense potential to track and identify individuals.

In some cases, these technologies are already operating in Australia without significant problems or widespread public objection. Facial recognition technologies are used in some Australian airports to aid check in, security and immigration processes to speed up processing and reduce costs while maintaining security [182,183]. However, there are significant privacy implications over the widespread use of facial recognition technology. There are extensive rules regarding earlier technologies that identify individuals, such as fingerprints, but in many respects the law has not caught up to technological capabilities such as facial recognition and the additional biometric information that is being collected, above and beyond fingerprints.

Microsoft in particular has been vocal in expressing concern over three key implications of the use of facial recognition technology. Microsoft President Brad Smith has stated [85]:

"First, especially in its current state of development, certain uses of facial recognition technology increase the risk of decisions and, more generally, outcomes that are biased and, in some cases, in violation of laws prohibiting discrimination.
Second, the widespread use of this technology can lead to new intrusions into people's privacy.
And third, the use of facial recognition technology by a government for mass surveillance can encroach on democratic freedoms."

These three uses of facial recognition technologies broadly encapsulate the challenges in rolling out this technology without adequate oversight and accountability mechanisms.

In response to the growing interest in these technologies there is a significant debate over how they should be used in Australia.

6.2.1 Case study: Surveillance technology in crisis situations

When used in service of humanitarian objectives, surveillance technologies such as facial and pattern recognition, geo-tracking and mapping can be a life-saving tool. In the event of a crisis there is an overwhelming amount of data that can be collected and analysed to aid in resolution of the situation, AI is well placed to manage this process.

A trial operation by police in India to test the use of facial recognition systems was able to scan the faces of 45,000 children in various children's homes and establish the identities of 2,930 children who had been registered as missing [184]. After bureaucratic difficulties between different agencies and the courts, the Delhi Police were able to utilise two datasets—60,000 children registered as missing, and 45,000 children residing in care institutions. From these two databases they were able to identify almost 3,000 matches [185].

Discussions are underway on how to use this system to identify missing children elsewhere in India. A key ingredient in this outcome was the ability for law enforcement to access these datasets [184].

While AI enabled surveillance may increase personal safety and reduce crime, the need to ensure that privacy is protected and that such technologies are not used to persecute groups is critical. Authorities need to give careful consideration to the use of AI in surveillance to ensure an appropriate balance is struck between protecting the safety of citizens and adopting intrusive surveillance measures that unfairly harm and persecute innocent people.

6.2.2 Monitoring employee behaviour with AI

Westpac Bank is among companies in Australia that are exploring the use of AI-enabled facial recognition technologies to monitor the moods of employees. Representatives have indicated that the goal is to "take the pulse" of teams across the organisation [186].

The use of facial recognition and mood detection AI to monitor employees can be used in ways that are ethical or unethical, and one way to assess this is to look at the goal of the exercise. Is it to benefit the welfare of employees, or is it to maximise profit? The Ethics Centre highlights the fact that AI technologies should keep the principle of "non-instrumentalism" in mind when designing technology [38]. This effectively means that humans should not merely become another part of the machine—the machine should serve people, not the other way around. In addition, the NHMRC National Statement on Ethical Conduct in Human Research states that "Respect for human beings involves giving due scope, throughout the research process, to the capacity of human beings to make their own decisions" [187] . When researching or utilising technologies that monitor people's emotions, it is important to ensure that their autonomy and right to make their own decisions are respected.

If, say, people's smiles were being logged by a machine and employees were disciplined for not being happy enough, and the goal was to put on a masquerade of happiness to please customers for profit reasons, then the machine is treating humans as another component of a profit-generating outcome. On the other hand, if the people's emotional state was assessed in order to deliver timely psychological assistance at the right time to people facing stress or an emotional breakdown, then the technology is serving people instead and could be defended on ethical grounds as long as it respected the autonomy of individuals and their right to choose not to participate.

6.2.3 Police and AI-enabled surveillance

The ability of facial recognition technologies to identify and track suspects means that increased police capabilities also need to come with commensurate oversight mechanisms. This is particularly important given that inaccuracies and inequities have been observed in the application of facial recognition technologies overseas.

In the United Kingdom, privacy advocates used Freedom of Information requests to gain access to the results of facial recognition programs by police. A report by Big Brother Watch indicated that in the London Metro area, police use of facial recognition proved 98% inaccurate, and no arrests were made. In South Wales, the technology was 91% inaccurate, and 15 arrests were made, roughly 0.005% of the number of people who were scanned with facial recognition technology. 31 innocent members of the public were asked to prove their identity [188]. The report pointed out that there is a lack of any real statutory guidance for the use of facial recognition technologies and warned of a potential "chilling effect" on people's attendance in public spaces if they know they are being observed through surveillance.

Academics commissioned by police to assess the use of facial recognition systems in South Wales found that it had helped in the arrests of around 100 suspects, but they stressed that it required police to adapt their operating methods to achieve results from the technology and this took time. They stated that at first, only 3% of matches proved accurate, but this improved to 46% over the course of the project [189]. They suggest that these technologies are best thought of as "assisted" facial recognition technologies and that a human is still required to confirm matches.

In Australia, government is considering the implications of facial recognition via the Identity Matching Services Bill 2018, which is still under discussion.

6.2.4 Balancing privacy with security

Groups such as the Human Rights Law Council have raised concerns over the ways in which personal biometric information may be shared between agencies [29]. They suggest that agencies should consider a framework put forward by the US-based Georgetown Law Center on Privacy and Technology which assesses the risks involved in police use of facial recognition technology [29,190]. This framework highlights five key risk factors to consider: [190]

1. Targeted versus dragnet searches (is the search just from convicted criminals and suspects, or innocent people too?)

2. Targeted versus dragnet databases (does the database include as many people as possible, including innocent people?)

3. Transparent versus invisible searches (do people know that their picture has been used in a search?)

6. Real time versus after the fact searches (is this search of past information, or tracking in real time?)

7. Established use versus novel use (how different is the application of this facial recognition compared to previous applications like fingerprinting?)

Australian authorities need to give careful consideration to the use of AI in surveillance and security, to ensure an appropriate balance is struck between protecting the safety of citizens and adopting intrusive surveillance measures. A privacy framework for law enforcement that incorporates the new capabilities delivered by facial recognition technologies could incorporate approaches like the Georgetown Framework, thus helping agencies ensure that facial recognition technologies are used in an appropriate manner.

6.3 Artificial Intelligence and Employment

In 2013 two University of Oxford academics, Carl Benedikt Frey and Michael Osborne, published a study [191] examining the impacts of automation on 702 unique occupation types in the US economy. They found 47% were at risk of being replaced. They also found a strong negative relationship between automation risks and wages (i.e. lower pay for jobs with a higher chance of being automated). This led to concern around the world about the possibilities of higher rates of unemployment and under-employment. The University of Oxford study was replicated in multiple jurisdictions. The Committee for the Economic Development of Australia (CEDA) commissioned a study [192] of the Australian economy and found a similar result with 44% of the workforce at risk to automation [193].

Recent research published by the United Kingdom Global Innovation Foundation Nesta [194] suggests that the original University of Oxford paper [191], and many others that have used a similar methodology, have overstated the job losses from automation. The Nesta report points out the AI will create many new jobs. It also notes that jobs impacted by AI don't just disappear; automation often just requires new skills and new tasks but the job stays intact. Accountants did not lose their jobs to spreadsheets; rather they learned how to use them and got better jobs. The coming decade of AI enablement will have a similar impact. A more recent meta-level study by the OECD [195] published in 2018 found that 14% of jobs have a "high risk" of automation and another 32% will be substantially changed.

Whilst there is much debate, and many other estimates (higher and lower), the weight of evidence suggests around half of all jobs will be significantly impacted (positively or negatively) by automation and digital technologies. A smaller, but still significant, number of jobs are likely to be fully automated requiring workers to transition into new jobs and new careers. Retraining, reskilling and strategic career moves can help people achieve better employment outcomes. A recent study by Google and consulting firm Alpha Beta [196] finds Australian workers will, on average, need to increase time spent learning new skills by 33% over their lifetime and that job tasks will change 18% per decade.

There is much that can be done by governments, companies and individuals to improve the chances of job retention and successful career transition in light of automation. One of the main issues is the importance of acting early; well before job loss occurs. An ethical approach to widespread AI-based automation of tasks performed by human workers requires helping the workers transition smoothly and proactively into new jobs and new careers.

6.4 Gender Diversity in AI workforces

Another aspect relating to employment ethics is associated with the gender balance within AI-technical workforces. Australia's Workplace Gender Equality Agency has indicated that the Professional, Scientific and Technical Services sector is only 40.9% female, and that full time female workers receive 23.7% less pay on average than their male counterparts. [197] When this is broken down into Computer System Design and related services, the proportion of women in the sector falls to 27% of employees in 2018 [197]. There is a risk that a lack of diversity in AI designers and developers results in a lack of diversity in the AI products they make. Many companies and research organisations in the technology sector are committed to addressing the gender imbalance.

The Government has recognised that Australia must have a deeper STEM talent pool and this is why it has supported the development of a Decadal Plan for Women in STEM to provide a roadmap for sustained increases in women's participation in STEM over the next decade.

The benefits of greater diversity in the ICT workforce will be felt across many dimensions of the Australian economy, including AI.

6.5 Artificial Intelligence and Indigenous Communities

Discussions and protocols that have focused on knowledge sharing and management between Indigenous people, science and decision-makers provide some valuable insights for AI frameworks and applications [198] in this context. This highlights three interrelated issues to consider:

1. AI based on data collected on and with Indigenous people needs to consider how data is collected and used so that it complies with Indigenous cultural protocols and human ethics and appropriately protects Indigenous intellectual property, knowledge and its use. As highlighted in a discussion paper commissioned by IP Australia and the Department of Industry, Innovation and Science, "Indigenous Knowledge is held for the benefit of a community or group as a whole and there can be strict protocols governing the use of Indigenous Knowledge, directed at gaining community approval" [199] . Guidelines for ethical research in Australian Indigenous studies offer a useful starting point to guide this effort [200].
2. Information is not intelligence and the analytical process by which Indigenous knowledge (is categorised, labelled, shared and incorporated into AI learning and feedbacks should be guided by cross-cultural collaborative approaches. The way in which indigenous knowledge is used has a direct bearing on the way it is collected—some uses of indigenous knowledge would not be considered acceptable to the communities they are drawn from, meaning that the uses would need to be clarified upfront.
3. AI needs to be open and accountable so that Indigenous people and organisations are clear about how AI learning is generated and why this information is used to inform decisions that affect Indigenous estates and lives.

The principles outlined in this document can provide some guidance on how to properly collect and handle indigenous knowledge, but are by no means the end point. Consideration of Net Benefits will need to place a strong emphasis not only on the application of the information, but the impacts on the communities that provide it. Further research into the relationship between AI and indigenous knowledge will be crucial in establishing proper standards and codes of conduct.

6.6 Key points

- Autonomous vehicles require hands-on safety governance and management from authorities, because systems will need to make choices on how to react under different circumstances and a system without a cohesive set of rules is likely to deliver worse outcomes that are not optimised for Australian road rules or conditions.
- AI-enabled surveillance technologies should consider "non-instrumentalism" as a key principle—does this technology treat human beings as one more cog in service of a goal, or is the goal to serve the best interests of human beings?
- In many ways, biometric data is replacing fingerprints as a key tool for identification as biometric data (which includes fingerprints) can now use other elements like facial recognition. The ease at which AI-enabled voice, face and gait recognition systems can identify people poses an enormous risk to privacy.
- AI technologies cannot be considered in isolation, they also need to take into account the context in which they will be used and the other technologies which will complement them.
- There are various factors that can be used to assess the risks of a facial recognition system. Designers of these systems should consider these factors.
- Workers and society will get better outcomes if we take proactive measures to assist smooth career transitions.
- There is a gender imbalance in terms of numbers and salaries in AI technical workforces which may need to be addressed.
- AI applied within indigenous communities needs to take into account cultural issues of importance.

7 A Proposed Ethics Framework

> As always in life people want a simple answer. It's like that lovely quote, for every complex problem in life there's always a simple answer and it's always wrong.
>
> Susan Greenfield

AI is a valuable tool to be harnessed, and one that can be used for many different goals. Already, companies and government agencies are finding that their increasing reliance on AI systems and automated decisions is creating ethical issues requiring resolution. With significant ramifications for the daily lives, fundamental human rights and economic prosperity of all Australians, a considered and timely response is required.

The eight core principles referred to throughout this report are used as ethical framework to guide organisations in the use or development of AI systems. These principles should be seen as goals that define whether an AI system is operating ethically. In each chapter of this report we highlighted specific principles that are associated with the case studies and discussions contained within them. It is important to note that all the principles should be considered throughout the design and use of an AI system not just those discussed in detail in each chapter.

1. **Generates net-benefits.** The AI system must generate benefits for people that are greater than the costs.
2. **Do no harm.** Civilian AI systems must not be designed to harm or deceive people and should be implemented in ways that minimise any negative outcomes.
3. **Regulatory and legal compliance.** The AI system must comply with all relevant international, Australian Local, State/Territory and Federal government obligations, regulations and laws.
4. **Privacy protection.** Any system, including AI systems, must ensure people's private data is protected and kept confidential plus prevent data breaches which could cause reputational, psychological, financial, professional or other types of harm to a person.
5. **Fairness.** The development or use of the AI system must not result in unfair discrimination against individuals, communities or groups. This requires particular attention to ensure the "training data" is free from bias or characteristics which may cause the algorithm to behave unfairly.
6. **Transparency and explainability.** People must be informed when an algorithm is being used that impacts them and they should be provided with information about what information the algorithm uses to make decisions.
7. **Contestability.** When an algorithm significantly impacts a person there must be an efficient process to allow that person to challenge the use or output of the algorithm.
8. **Accountability.** People and organisations responsible for the creation and implementation of AI algorithms should be identifiable and accountable for the impacts of that algorithm.

7.1 Putting principles into practice

The principles provide goals to work towards, but goals alone are not enough. The remainder of this section will explore the ways in which individuals, teams and organisations can reach these goals. To support the practical application of the core ethical principles, an AI toolkit has been referenced throughout the report as potential instruments of action.

The toolkit does not address all potential solutions regarding the governance of AI in Australia and is intended to provide a platform upon which to build knowledge and expertise around the ethical use and development of AI in Australia. It is unlikely that there will be a one size fits all approach to address the ethical issues associated with AI [201]. In addition the approaches taken to address these issues are unlikely to remain static over time.

This chapter provides guidance for individuals or teams responsible for any aspect of the design, development and deployment of any AI-based system that interfaces with humans. It can help AI practitioners address three important questions:

- What is the purpose of the AI system?
- What are the relevant principles to guide the ethical use and application of the AI system?
- How do you assess the requirements of meeting these ethical principles?

What are the tools and processes that can be employed to ensure the AI system is designed, implemented and deployed in an ethical manner? Additionally, there is a sample risk framework which can guide AI governance teams in assessing the levels of risk in an AI system.

We would invite stakeholders as part of the public consultation to share their thoughts and expertise on how ethical AI can be practically implemented.

7.1.1 Impact assessments

These are auditable assessments of the potential direct and indirect impacts of AI, which address the potential negative impacts on individuals, communities and groups, along with mitigation procedures.

Algorithmic impact assessments (AIA) are designed to assess the potential impact that an AI system will have on the public. They are often used to assess automated decision systems used by governments [26,202]. The AI Now Institute have developed an AIA and are urging the recently appointed New York City (NYC) task-force to consider using their framework to ensure that all automated decision systems used by the NYC government are made according to principles of equity, fairness and accountability [2,26].

The four key goals of the AI Now Institute’s AIA are:

- “Respect the public’s right to know which systems impact their lives by publicly listing and describing automated decision systems that significantly affect individuals and communities
- Increase public agencies’ internal expertise and capacity to evaluate the systems they build or procure, so they can anticipate issues that might raise concerns, such as disparate impacts or due process violations
- Ensure greater accountability of automated decision systems by providing a meaningful and ongoing opportunity for external researchers to review, audit, and assess these systems using methods that allow them to identify and detect problems
- Ensure that the public has a meaningful opportunity to respond to and, if necessary, dispute the use of a given system or an agency’s approach to algorithmic accountability.”

In addition to assessing algorithms, impact assessments can be designed to address other important ethical issues associated with AI. The Office of the Australian Information Commissioner (OAIC) has developed a privacy impact assessment to identify the impact that a project could have on individual privacy [203]. There is also an affiliated eLearning guide, to help provide guidance to organisations about 'privacy by design' approaches to data use [204].

The adoption and use of standard, auditable, impact assessments by organisations developing or using AI in Australia would help encourage accountability and ensure that ethical principles are considered and addressed before AI was implemented.

7.1.2 Review processes

Specialised professionals or groups can review AI and/or use of AI systems to ensure that they adhere to ethical principles and Australian policies and legislation.

In many cases, Australia is likely to be importing "off the shelf" AI developed internationally under different regulatory frameworks. In these cases adequate review process will be key to ensuring that the technology meets Australian standards and adheres to ethical principles.

Alternatively, in some cases it may be permissible to use AI programs to review other AI systems. Several companies have developed tools to that are able to effectively assess algorithms used by AIs and report on how the system is operating and whether it is acting fairly or with bias [27]. IBM has released an open source, cloud based software that creates an easy to use visual representation that shows how the algorithms are generating decisions [205]. In addition, it can assess the algorithm's accuracy, fairness and performance. Microsoft and Google are working on similar tools to assess algorithms for bias [27]. The use of such technologies could improve our ability to efficiently, effectively and objectively review the components of AI to ensure that they adhere to key ethical principles. However, if utilised, these are AI enabled technologies would require a significant degree of scrutiny to ensure that they did not have the same flaws that they were purporting to assess.

7.1.3 Risk assessments

The assessment of AI is largely an exercise in accounting for and addressing risks posed by the use of the technology [201]. As such, consideration should be given to whether certain uses of AI require additional assessment, these may be considered to be threshold assessments. FATML have developed a Social Impact Statement that details requirements of developers of AI to consider who will be impacted by the algorithm and who is responsible for that impact [206]. Similar assessments may be well placed to identify high risk applications and uses of AI that require additional monitoring or review.

There are additional potential risks when AI is used in vulnerable populations and minorities. In these cases we should consider whether additional scrutiny is required to ensure it is fair. For example, when conducting research involving human participants additional considerations must be made when dealing with vulnerable groups and minorities [207].

One argument against this concept of risk based levels of assessment is that the standard level of assessment should ensure that AI across all spectrums is acting and used according to the key ethical principles. Perhaps we should expect that a standard prescribed course of action should be rigorous enough to ensure that low to high risk AI adheres to core ethical principles.

7.1.4 Best practice guidelines

This involves the development of accessible cross industry best practice principles to help guide developers and AI users on gold standard practices.

Best practice guidelines encompass the best available evidence and information to inform practice. For example, The Office of Australia's Fair Work Ombudsman has published various best practice guides for employers and employees to help identify and implement best practice initiatives into their workplaces [208]. Similar guidelines could be developed to provide best practice initiatives that support the ethical use and development of AI. The use of these adaptable and flexible best practice guides rather than rigid policies fit well with the dynamic nature of AI and the difficulty in predicting what is coming next [209]. It would be straightforward to adjust best practice guidelines as situations and scenarios change over time.

The Australian government has already developed a best practice guide to provide strategies and information about the best practice use of technology to make automated decisions by their agencies [210]. Similar guidelines could be developed and promoted to support consideration of the core ethical issues associated with AI use and development.

7.1.5 Education, training and standards

Standards and certification of AI systems is being actively explored both nationally and internationally.

In Australia, the provision and certification of standards are generally overseen by relevant industry bodies. Doctors are accountable to medical bodies and there are extensive regulations on their behaviour, and the Australian Medical Association has a code of ethics for guidance [211]. Electricians, plumbers and people involved in air-conditioning repair all require certification to demonstrate their skills and guarantee public safety [212]. States have various requirements regarding certification for repairing motor vehicles [213]. Engineers Australia provides accreditation for programs that train engineers in coordination with international standards [214]. There are industry bodies such as Data Governance Australia which are examining data principles [215].

One area where the implementation of standards could have a large positive impact on ethical AI in Australia relates to data scientists. Currently, there is no agreed upon accreditation or standards that govern data science as a profession. Designers of algorithms which may have significant impacts on public well-being are operating within a profession with relatively limited guidance or oversight.

Australia's national standards body, Standards Australia is working with industry stakeholders in developing an AI Standards roadmap to guide the development of an Australian position on AI standards.

Dr Alan Finkel (Australia's Chief Scientist), has also proposed a framework for voluntary certification of ethical AI by qualified experts which his office is currently exploring further.

Internationally the International Standards Organisation (ISO) has a technical committee, which Australia is an observer, developing standards on AI ((ISO) (ISO/IEC JTC 1/SC 42 – Artificial Intelligence). This includes both technical and ethical standards.

There is significant scope within Australia to provide more guidance for the formulation of standards to govern designers of AI systems.

7.1.6 Business and academic collaboration in Australia

A key focus of Australia's National Innovation and Science Agenda is the promotion of collaboration through, "funding incentives so that more university funding is allocated to research that is done in partnership with industry; and invest over the long term in critical, world-leading research infrastructure to ensure our researchers have access to the infrastructure they need" [216]. One such initiative is the provision of an Intellectual Property (IP) toolkit to help resolve complex issues that can arise over IP when industry and academics collaborate [217].

The Australian Technology Network is a partnership between several innovative universities committed to developing collaboration with industry. In a recent report they put forward five recommendations to foster these relationships [218]:

- "Expand in place supporting structures to deepen PhD and university collaboration with industry
- Ensure initiatives targeting PhD employability have broad scale
- Link a portion of PhD scholarships to industry collaboration
- Implement a national communication strategy to improve awareness and develop a deeper understanding in industry of the PhD"

Quality research into addressing ethical AI by design and implementation is key to ensuring that Australia stays ahead of the curve. Without methods of accessible transfer of knowledge from theory to practice the impact is lost. Collaboration is increasingly important between researchers and the tech industry to ensure that AI is developed and used ethically and should be prioritised.

7.1.7 Monitoring AI

This consists of regular monitoring of AI or automated decision systems for accuracy, fairness and suitability for the task at hand. This should also involve consideration of whether the original goals of the algorithm are still relevant.

Promotion of regular assessment of AI systems and how they are being used will be a key tool to ensure that all of the core ethical principles are being addressed. Although initial assessments before the deployment of AI are critical they are unlikely to provide the scope needed to assess the ongoing impact of the AI in the changing world.

Regular monitoring of AI to assess whether it is still suitable for the task at hand and whether it still adheres to the core ethical principles could be encouraged in best practice guidelines or as part of ongoing impact assessments.

7.1.8 Recourse mechanisms

Are there avenues for appeal when an automated decision or the use of AI negatively affects a member of the public?

The GDPR and the UK's Data Protection Act both include the requirement for individuals to be informed about the use of automated decisions that affect them and provide the opportunity to contest those findings [92]. Recourse mechanisms help promote transparency between organisations using automated decisions and the users affected by the systems. They also engender trust between individuals and organisations and could be used to improve public acceptance of the use of AI. The provision of recourse mechanisms has become increasingly important in cases of black box algorithms where the process of how the system came to a decision or judgement cannot be elucidated.

It may be important to consider that there may be additional complexities associated with the provision of recourse mechanisms. In situations where the AI systems were found to be faulty, demands could be made for compensation if any damages were incurred as a result of the impact of the AI system on the individual. This ties into the principle of suitability of AI systems and the need to ensure that they are appropriate for the task at hand and can perform in a manner that does not cause unacceptable levels of harm when weighed against the benefits of their use.

7.1.9 Consultation

Without public support there is no destination for the momentum which AI is building. Investing in avenues for public feedback and dialogue on AI will be key to ensuring that the development and use of AI is in line with what Australians want. This tool is related to the principle of net benefit and the need for AI systems to generate benefits greater than the costs. As discussed at the 2018, Global Symposium for Regulators in regards to public consultation, “Keep an open door and an open mind. When it comes to AI, no one understands all of the problems, let alone all of the solutions. Hearing from as many perspectives as possible will expose policymakers and regulators to issues that may not have been on their radar and creative solutions they may not have tried otherwise. And some of these solutions may not require law or regulation” [219] .

Regular large scale consultation with various stakeholders including the general public, academics and industry members is of critical importance when developing AI regulations [209]. A diverse range of inputs can only be collected from a diverse group of stakeholders. Various organisations developing materials that provide information about ethics and AI have included a lengthy consultation process and courted input from diverse and varied sources [36,72,73]. Consultation is a valuable tool that can help to better understand the spectrum of ideas, concerns and solutions regarding ethical AI.

7.2 Example Risk Assessment Framework for AI Systems

Risk assessments are commonly used to assess risk factors that have the potential to cause harm. The assessment of AI is largely an exercise in accounting for and addressing risks posed by the use of the AI system. They are also useful to provide a threshold and triggers for additional action and risk mitigation processes. This preliminary guide is for individuals or teams responsible for any aspect of the design, development and deployment of any AI-based system that interfaces with humans. Its purpose is to guide AI practitioners to address three questions: What is the purpose of the AI system? What are the relevant principles to guide the ethical use and application of the AI system? What are the tools and processes that can be employed to ensure the AI system is designed, implemented and deployed in an ethical manner?

This is just one example framework, which cannot stand in for frameworks tailored for each individual application of AI.

The first table examines the probability of risk, together with the consequence. When a risk has both a high probability of occurring and more negative outcomes, the consequences become more severe.

Likelihood of risk	**Consequence**				
	Insignificant risk	**Minor risk**	**Moderate risk**	**Major risk**	**Critical risk**
Rare	Low	Low	Moderate	High	High
Unlikely	Low	Moderate	Moderate	High	Extreme
Possible	Low	Moderate	High	High	Extreme
Likely	Moderate	High	High	Extreme	Extreme
Almost certain	Moderate	High	High	Extreme	Extreme

The second table examines the factors that can cause an AI application to contain more risk. The rows near the top carry relatively little risk, while the rows near the bottom contain more risk. Different scenarios may contain more or less risk depending on the individual circumstances, but this guide provides a general overview of areas likely to contain risks which ought to be considered before implementation. It is also worth noting that although there is a column for the number of people affected, severe repercussions for just a single person would still be viewed as a major or critical consequence.

Privacy Protection /Consequence	**Fairness /Consequence**	**Physical harm /Consequence**	**Contestability /Consequence**	**Accountability /Consequence**	**Regulatory and legal Compliance /Likelihood**	**Transparency and explainability /Likelihood**	**Number of people affected**
Insignificant – no private or sensitive data is used by the AI	**Insignificant** – has no effect on a person's human rights	**Insignificant** – application cannot control or influence other systems	**Insignificant –** the application operates on an opt-in basis and no intervention is required to reverse outcome in the event someone decides to opt out	**Insignificant –** clear accountability of outcomes	**Insignificant** – consent gained for use of data. Application operates in familiar legal territory.	**Insignificant** – inputs, algorithm and output are bounded and well understood	**Insignificant** – application will not affect individuals
Minor – uses small number of people's private data	**Minor** – clear opt-in and opt-out of application	**Minor** – application controls equipment incapable of causing significant harm, public nuisance or surveillance	**Minor** – application automatically opts people in, but there is clear notification and it is easy to opt out. It is easy to obtain human assistance to do so.	**Minor** – there is legal precedent for similar outcomes and a clear chain of responsibility	**Minor** – an identical or similar application of AI has legal precedent demonstrating compliance. Consent has clearly been gained.	**Minor** – the application uses difficult-to-explain AI like neural nets, but the inputs are clear and there are no cases of totally unexpected, inexplicable outputs	**Minor** –the application will run only within an organisation and affect a small number of people
Major – uses a large number of people's private data	**Major** – opt-in but opt-out of application is unclear	**Minor** – application may control heavy equipment or hazardous material but there is very limited potential to harm people or cause public nuisance	**Major** – application will be used widely among the public but there are little or no human resources available for assistance or appeal	**Minor** – there is reasonably clear delineation of accountability between users and developers	**Minor** – there is little legal precedent for the application, but extensive legal advice has been sought and the application has been reviewed by third parties	**Minor:** the application has unexpected outputs but they are periodically reviewed until understood. External review and collaboration is fostered	**Minor**– the algorithm will affect a small community of people who are well-informed about its use
Major – application is designed in a way that makes it likely to gather information on individuals without their express consent	**Major** – person has no choice to opt-out of application and AI has an effect on a single person's human rights	**Major** – application controls heavy equipment / hazardous material and is expected to operate in a public space	**Major** – there is no easy way to opt out	**Major** – there is little or no legal precedent for this application and there is no separation of accountability between users and developers	**Major** – there is no legal precedent for the application and no third parties or legal experts have been consulted. Consent is unclear.	**Major:** the application's outputs are inexplicable. Review is of limited effectiveness in understanding them	**Major:** the application will affect a large number of people around the country
Critical – uses a major database of private or sensitive (e.g. health) data	**Critical** – has an effect on a large population's human rights	**Critical** – application can control equipment that could cause loss of life, or equipment is designed to secretly gather personal information	**Critical –** outcomes of the application not opt-out and person affected has no recourse to change the outcome	**Critical –** unclear legal accountability for outcomes	**Critical** – no consent gained for use of large quantities of private data. There is no clear legal precedent.	**Critical** – inputs are uncontrolled, the algorithm is not well-understood and the outputs are not understood	**Critical** – the application affects a national or global audience of people

There are a variety of actions that can be taken to mitigate risk. These are just some of many that have been explored in this report. These measures are by no means comprehensive—consultation with external organisations will yield additional solutions in many cases. Additional measures may also emerge with new research, technologies and practices.

Risk	**Actions**
Low	• Internal Monitoring • Testing • Review industry standards
Moderate	• Internal Monitoring • Consider how to lower risk • Risk mitigation plan • Internal review • Testing • Impact assessments • External review
High	• Internal/External Monitoring • Consider how to lower risk • Risk mitigation plan • Impact assessments • Internal and external review • Testing • Consultation with specialists • Detailed appeals/opt out plan • Additional human resources to handle inquiries/appeals • Legal advice sought • Liaise with industry partners, government bodies on best practice
Extreme	• Unacceptable risk

We invite stakeholders as part of the public consultation to share their thoughts and expertise on how ethical AI can be practically implemented.

8 Conclusion

“Humans are allergic to change. They love to say, "We've always done it this way." I try to fight that. That's why I have a clock on my wall that runs counter-clockwise."

Grace Hopper

This framework discussion paper is intended to guide Australia’s first steps in the journey towards integrating policies and strategies to provide a landscape that supports the positive development and use of AI.

The principles and toolkit items provide practical, accessible approaches to harness the best that AI can offer Australia, while addressing the risks. AI is an opportunity; one that has the potential to provide a better future with fairer processes and tools to address important environmental and social issues.

However, reaching this future will require input from stakeholders across government, business, academia and broader society. AI developers cannot be expected to bear the responsibility for achieving these outcomes all on their own. Further collaboration will be of the utmost importance in reaching these goals.

9 References

1 Ethics Commission Federal Ministry of Transport and Digital Infrastructure. 2017. Automated and connected driving. German Government. Germany.

2 New York City Hall. 2018. Mayor de Blasio announces first-in-nation Task Force to examine automated decision systems used by the city. Mayor's Office.

3 House of Lords Select Committee on Artificial Intelligence. 2018. AI in the UK: Ready, willing and able? UK Parliament. United Kingdom.

4 European Commission. 2018. Artificial Intelligence for Europe.

5 Canadian Institute for Advanced Research. 2017. Pan-Canadian artificial intelligence strategy.

6 Australian Government. 2018. 2018-2019 Budget overview. Canberra.

7 National Institute for Transforming India. 2018. National strategy for artificial intelligence.

8 Villani C. 2018. For a Meaningful Artificial Intelligence.

9 Rosemain M and Rose M. France to spend $1.8 billion on AI to compete with U.S., China. Reuters. 29 March 2018.

10 The Japanese Society for Artificial Intelligence. 2017. The Japanese society for artificial intelligence ethical guidelines. Japan.

11 European Commission. 2018. High-level expert group on artificial intelligence.

12 UK Parliament. 2018. World first Centre for Data Ethics and Innovation: Government statement.

13 Infocomm Media Development Authority. 2018. Composition of the Advisory Council on the Ethical Use of Artificial Intelligence and Data. Minister for Communications and Information. Singapore.

14 The State Council: The People's Republic of China. 2018. Guidelines to ensure safe self-driving vehicle tests. The State Council: The People's Republic of China. The People's Republic of China.

15 Senate Community Affairs Committee Secretariat. 2017. Design, scope, cost-benefit analysis, contracts awarded and implementation associated with the Better Management of the Social Welfare System initiative. Parliament of Australia.

16 Angwin J, Larson J, Mattu S et al. 2016. Machine bias risk assessments in criminal sentencing. ProPublica.

17 Dieterich W, Mendoza C, Brennan T. 2016. COMPAS risk scales: Demonstrating accuracy and predictive parity. Northpointe:

18 Angwin J, Larson J. 2016. ProPublica responds to company's critique of Machine Bias story. ProPublica.

19 Angwin J, Larson J. 2016. Bias in criminal risk scores is mathematically inevitable, researchers say. ProPublica.

20 Office of the Australian Information Commissioner. 2018. Publication of Medicare Benefits Schedule and Pharmaceutical Benefits Schedule data: Commissioner initiated investigation report. Australian Government.

21 Australian Government. International human rights system. Attorney General's Department.

22 Australian Government. 1988. Privacy Act 1988, No. 119, 1988 as amended. Australian Government. Canberra.

23 Corbett-Davies S, Pierson E, Feller A et al. 2016. A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear. The Washington Post.

24 Moses L B, Chan J. 2016. Algorithmic prediction in policing: assumptions, evaluation, and accountability. Policing and Society, 28(7): 806-822.

25 Australian Government. 1999. Social Security (Administration) Act 1999: No 191, 1999. Federal Register of Legislation.

26 Reisman D, Schultz J, Crawford K et al. 2018. Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability.

27 Kleinman Z. IBM launches tool aimed at detecting AI bias.

28 Khadem N. Tax office computer says yes, Federal Court says no. ABC. 8 October 2018.

29 Human Rights Law Centre. 2018. The dangers of unregulated biometric use: Submission to the Inquiry into the Identity-matching Services Bill 2018 and the Australian Passports Amendment (Identity-matching Services) Bill 2018. Human Rights Law Centre. Australia.

30 Awad E, Dsouza S, Kim R et al. 2018. The Moral Machine experiment. Nature, 563(7729): 59-64.

31 Berg A, Buffie E, Zanna L-F. 2018. Should we fear the robot revolution? (The correct answer is yes). International Monetary Fund.

32 CSIRO. 2015. Robots to ResQu our rainforests. CSIRO.

33 Ray S. 2018. Data guru living with ALS modernizes industries by typing with his eyes. Microsoft News.

34 International Electrotechnical Commission. Functional Safety and IEC 61508.

35 Redrup Y. 2018. Google to make AI accessible to all businesses with Cloud AutoML. Australian Financial Review.

36 Australian Human Rights Commission. 2018. Human rights and technology issues paper. Sydney.

37 McCarthy J, Minsky M L, Rochester N et al. 1955. A Proposal for the Dartmouth Summer Research Project on artificial intelligence.

38 Beard M, Longstaff S. 2018. Ethical by design: principles for good technology. The Ethics Centre. Sydney, Australia.

39 Kranzberg M. 1986. Technology and History: "Kranzberg's Laws". Technology and Culture, 27(3): 544-560.

40 Dutton T. An Overview of National AI Strategies. Medium [Internet]. 2018 Available from: https://medium.com/politics-ai/an-overview-of-national-ai-strategies-2a70ec6edfd

41 British Broadcasting Corporation. 2017. Google DeepMind NHS app test broke UK privacy law. United Kingdom.

42 Elvery S. 2017. How algorithms make important government decisions — and how that affects you. ABC. Australia.

43 Australian Government. 2007. Automated Assistance in Administrative Decision-Making.

44 UN. 1966. International Covenant on Civil and Political Rights. United Nations.

45 UN. 1966. International Covenant on Economic, Social and Cultural Rights. United Nations.

46 UN. 1966. International Convention on the Elimination of all Forms of Racial Discrimination. United Nations.

47 UN. 1979. Convention on the Elimination of All Forms of Discrimination against Women. United Nations.

48 UN. 1984. Convention against Torture and other Cruel, Inhuman or Degrading Treatment or Punishment. United Nations.

49 UN. 1989. Convention on the Rights of the Child. United Nations.

50 United Nations. 2007. Convention on the rights of persons with disabilities. United Nations.

51 United Nations. 1948. The universal declaration of human rights. United Nations.

52 Australian Government. 2011. Human rights (Parliamentary Scrutiny) act 2011, No 186, 2011. Federal Register of Legislation.

53 Australian Government. Statements of compatibility. Attorney General's Department. Australia.

54 Attorney-General's Department. Australia's anti-discrimination law. Australian Government. Australia.

55 Cossins D. 2018. Discriminating algorithms: 5 times AI showed prejudice. New Scientist.

56 Australian Government - Productivity Commission. 2017. Data availability and use, productivity commission inquiry report. 82):

57 Department of the Prime Minister and Cabinet. 2018. New Australian Government data sharing and release legislation: Issues paper for consultation. Australian Government. Canberra.

58 Office of the Victorian Information Commissioner. 2018. Artificial intelligence and privacy. Office of the Victorian Information Commissioner. Victoria.

59 Australian Government. Privacy Act 1988. Federal Register of Legislation.

60 Australian Government. Australian Privacy Principles. Office of the Australian Information Commissioner.

61 Australian Government. Credit Reporting. Office of the Australian Information Commissioner.

62 Australian Government. Tax File Numbers. Office of the Australian Information Commissioner.

63 Australian Government. Health Information and Medical Research. Office of the Australian Information Commissioner.

64 European Union. Art. 22 GDPR Automated individual decision-making, including profiling.

65 Wachter S, Mittlestadt B, Floridi L. 2017. Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation. International Data Privacy Law, 7(2): 76-99.

66 European Group on Ethics in Science and New Technologies. 2018. Statement on artificial intelligence, robotics and 'autonomous' systems. European Commission. Brussels.

67 European Commission, High Level Expert Group on AI. 2018. Draft Ethics guidelines for trustworthy AI.

68 Villani C. 2018. For a meaningful artificial intelligence: Towards a French and European strategy. French Government.

69 Belot H, Piper G, Kesper A. 2018. You decide: Would you let a car determine who dies?

70 Whittlestone J, Nyrup R, Alexandrova A et al. 2019. Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research.

71 IEEE. The IEEE global initiative on ethics of autonomous and intelligent systems. IEEE Standards Association.

72 The IEEE global initiative on ethics of autonomous and intelligent systems. 2016. Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems.

73 The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. 2017. Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems.

74 Campolo A, Sanfilippo M, Whittaker M et al. 2017. AI now 2017 report. AI Now Institute.

75 Whittaker M, Crawford K, Dobbe R et al. 2018. AI Now Report 2018.

76 Stone P, Brooks R, Brynjolfsson E et al. 2016. One hundred year study on artificial intelligence: Report of the 2015-2016 study panel. Stanford University. Stanford, USA.

77 Future of Life Institute. 2017. Asilomar AI principles.

78 The Public Voice. 2018. Universal Guidelines for Artificial Intelligence. Electronic Privacy Information Center. Brussels, Belgium.

79 Partnership on AI. Partnership on AI web page. Available from: https://www.partnershiponai.org/partners/.

80 Pichai S. 2018. AI at Google: Our principles. Google.

81 Specktor B. 2018. Google will end its 'evil' partnership with the US Military, but not until 2019. Live Science.

82 Newcomer E. 2018. What Google's AI principles left out. Bloomberg. U.S.A.

83 Greene T. 2018. Google's principles for developing AI aren't good enough. The Next Web.

84 Hern A. 2017. Whatever happened to the DeepMind AI ethics board Google promised?

85 Smith B. 2018. Facial recognition: It's time for action. Microsoft On the Issues - the Official Microsoft Blog

86 Microsoft. Our Approach to AI (website).

87 IBM. 2018. Everyday Ethics for Artificial Intelligence.

88 Dafoe A. 2018. AI governance: A research agenda. Future of Humanity Institute. Oxford.

89 Bostrom N, Yudkowsky E. 'The ethics of artificial intelligence' In: The Cambridge Handbook of Artificial Intelligence. Cambridge University Press; 2014.

90 The AI Initiative. The AI initiative recommendations. The Future Society. Harvard Kennedy School.

91 Nguyen P, Solomon L. 2018. Emerging issues in data collection, use and sharing. Consumer Policy Research Centre. Australia.

92 European Commission. European Commission GDPR Home Page. [9 August 2018]. Available from: https://ec.europa.eu/commission/priorities/justice-and-fundamental-rights/data-protection/2018-reform-eu-data-protection-rules_en.

93 UK Government. 2018. Data Protection Act.

94 Rosenberg M, Confessore N, Cadwalladr C. 2018. How Trump Consultants Exploited the Facebook Data of Millions. The New York Times

95 Vengattil M. 2018. Cambridge Analytica begins insolvency proceedings in the UK. Financial Review.

96 Bhardwaj P. 2018. Eight weeks after the Cambridge Analytica scandal, Facebook's stock price bounces back to where it was before the controversy. Business Insider Australia. Australia.

97 Government A. 2017. Privacy Amendment (Notifiable Data Breaches) Act 2017.

98 Office of the Australian Information Commissioner. 2018. Notifiable data breaches quarterly statistics report, 1 April-30 June 2018. Australian Government.

99 Institute P. 2017. Ponemon Institute's 2017 Cost of Data Breach Study: Australia.

100 United States Government Accountability Office. 2018. Data protection - actions taken by Equifax and federal agencies in response to the 2017 breach. U.S. Government.

101 Newman L H. 2017. Equifax officially has no excuse. Wired.

102 Fleishman G. 2018. Equifax data breach, one year later: Obvious errors and no real changes, new report says. Fortune.

103 Australian Government. 2015. Australian Government public data policy statement.

104 Department of the Prime Minister and Cabinet. 2018. The Australian Government's response to the productivity commission data availability and use inquiry. Australian Government. Canberra.

105 The Treasury. 2018. Consumer data right. Australia.

106 Open Knowledge Network. 2016. Global Open Data Index, Place Overview.

107 Hauge M V, Stevenson M D, Rossmo D K et al. 2016. Tagging Banksy: using geographic profiling to investigate a modern art mystery. Journal of Spatial Science, 61(1): 185-190.

108 Balthazar P, Harri P, Prater A et al. 2018. Protecting your patients' interests in the era of big data, artificial intelligence, and predictive analytics. Journal of the American College of Radiology, 15(3, Part B): 580-586.

109 Harvard Business School Digital Initiative. 2018. Tay: Crowdsourcing a PR nightmare. Harvard Business School. U.S.A.

110 Calmon F, Wei D, Vinzamuri B et al. 2017. Optimized pre-processing for discrimination prevention. Neural Information Processing Systems Conference

111 Rahwan I. 2018. Society-in-the-loop: Programming the algorithmic social contract. Ethics and Information Technology, 20(1): 5-14.

112 Gunning D. 2018. Explainable artificial intelligence (XAI). Defense Advanced Research Projects Agency.

113 Langford C. 2017. Houston schools must face teacher evaluation lawsuit. Courthouse News Service.

114 Parasuraman R, Riley V. 1997. Humans and automation: Use, misuse, disuse, abuse. Human Factors, 39(2): 230-253.

115 Skitka L J, Mosier K L, Burdick M. 1999. Does automation bias decision-making? International Journal of Human-Computer Studies, 51(5): 991-1006.

116 National Transportation Safety Board. Enbridge Incorporated Hazardous Liquid Pipeline Rupture and Release.

117 Wesley D, Dau L A. 2017. Complacency and Automation Bias in the Enbridge Pipeline Disaster. Ergonomics in Design, 25(1): 17-22.

118 Perry M. 2014. iDecide: The legal implications of automated decision-making. Cambridge Public Law Conference

119 National Transportation Safety Board. 2018. Preliminary report highway 18mh010.

120 Smith B W. 2017. Automated Driving and Product Liability. Michigan State Law Review, 1(

121 Burns K. 2016. Judges, 'common sense' and judicial cognition. Griffith Law Review, 25(3): 319-351.

122 Danziger S, Levav J, Avnaim-Pesso L. 2011. Extraneous factors in judicial decisions. PNAS, 108(17): 6889-6892.

123 Corbyn Z. Hungry judges dispense rough justice. Nature [Internet]. 2011 Available from: https://www.nature.com/news/2011/110411/full/news.2011.227.html.

124 Weinshall-Margel K, Shapard J. 2011. Overlooked factors in the analysis of parole decisions. PNAS, 108(42):

125 Greenwood A. 2018. Speaking remarks: The art of decision-making. Federal Court of Australia. Digital Law Library.

126 Australian Human Rights Commission. 2014. A quick guide to Australian discrimination laws.

127 World Health Organisation. Skin cancers: Who is most at risk of getting skin cancer?

128 Angwin J, Larson J. 2015. The Tiger Mom tax: Asians are nearly twice as likely to get a higher price from Princeton Review. ProPublica.

129 Stobbs N, Hunter D, Bagaric M. 2017. Can sentencing be enhanced by the use of artificial intelligence? Criminal Law Journal, 41(5): 261-227.

130 Australian Federal Police. 2017. Policing for a safer Australia: Strategy for future capability. Australian Federal Police.

131 Supreme Court of the United States Blog. 2017. Loomis v. Wisconsin: Petition for certiorari denied on June 26, 2017.

132 Carlson A M. 2017. The Need for Transparency in the Age of Predictive Sentencing Algorithms. Iowa Law Review, 103(1):

133 Lohr S. Facial Recognition Is Accurate, if You're a White Guy. New York Times.

134 Dastin J. Amazon scraps secret AI recruiting tool that showed bias against women. Reuters. 10 October 2018.

135 ANZ. 2018. Organisations are turning to Artificial Intelligence to improve their recruitment processes. Here's how it can benefit candidates. efinancialcareers.

136 Stanley J. 2018. New Orleans Program Offers Lessons In Pitfalls Of Predictive Policing. ACLU.

137 Delaney J. France, China, and the EU all have an AI strategy. Shouldn't the US? Wired.com.

138 Lapowsky I. How the LAPD uses data to predict crime. WIRED. 22 May 2018.

139 Crockford T. App data predicts when, where Brisbane criminals will strike next. Sydney Morning Herald. 31 October 2018.

140 Shakila Khan Rumi K D, Flora Dilys Salim. 2018. Crime event prediction with dynamic features. EPJ Data Science, 7(43):

141 Chen S. Doctors said the coma patients would never wake. AI said they would - and they did. South China Morning Post. 8 September 2018.

142 Federal Register of Legislation. 2007. Private Health Insurance Act 2007: Compliation No. 29.

143 André Q, Carmon Z, Wertenbroch K et al. 2017. Consumer choice and autonomy in the age of artificial Intelligence and big data. Customer Needs and Solutions, 5(1-2): 28-37.

144 Bloomberg Government. 2018. Transcript of Mark Zuckerberg's senate hearing. The Washington Post.

145 Hern A. 2018. Facebook ad feature claims to predict user's future behaviour. The Guardian.

146 Kramer A D I, Guillory J E, Hancock J T. 2014. Experimental evidence of massive-scale emotional contagion through social networks. PNAS, 111(24): 8788.

147 Kim H, Garrido P, Tewari A et al. 2018. Deep video portraits. Siggraph 2018

148 Just V. 2018. AI could make dodgy lip sync dubbing a thing of the past. University of Bath. United Kingdom.

149 European Commission. 2016. The EU internet handbook: Cookies. European Commission.

150 Knight W. 2018. The Defense Department has produced the first tools for catching deepfakes. MIT Technology Review. USA.

151 Society of Automotive Engineers. 2018. Taxonomy and definitions for terms related to on-road motor vehicle automated driving systems: J3016_201806.

152 National Transport Commission. 2018. Automated vehicles in Australia.

153 National Transport Commission. 2017. Guidelines for Trials of Automated Vehicles in Australia.

154 National Transport Commission. 2017. Clarifying control of automated vehicles: Policy paper.

155 National Transport Commission. 2018. Changing driving laws to support automated vehicles: Policy paper.

156 Johnston M. 2018. Federal govt unveils future transport tech office. ITnews.

157 Bloomberg Philanthropies. 2017. Taming the autonomous vehicle: A primer for cities. Aspen Institute Center for Urban Innovation. USA.

158 Ackerman E. 2017. Toyota's Gill Pratt on self-driving cars and the reality of full autonomy. IEEE Spectrum.

159 Mervis J. 2017. Are we going too fast on driverless cars? Science.

160 Marowits R. 2017. Self-driving Ubers could still be many years away, says research head. CTV News. Canada.

161 Truett R. 2016. Don't worry: Autonomous cars aren't coming tomorrow (or next year). AutoWeek.

162 Litman T. 2018. Autonomous vehicle implementation predictions implications for transport planning. Victoria Transport Policy Intitute. Victoria.

163 Stern R E, Cui S, Delle Monache M L et al. 2018. Dissipation of stop-and-go waves via control of autonomous vehicles: Field experiments. Transportation Research Part C: Emerging Technologies, 89(205-221.

164 Fagnant D J, Kockelman K. 2015. Preparing a Nation for Autonomous Vehicles: Opportunities, Barriers and Policy Recommendations for Capitalizing on Self-Driven Vehicles. Transportation Research167-181.

165 International Transport Forum. 2015. Big data and transport: Understanding and assessing options. OECD Publishing. Paris, France.

166 Schoettle B, Sivak M. 2015. Potential impact of self-driving vehicles on household vehicle demand and usage. The University of Michigan Transportation Research Institute. Michigan, USA.

167 Trommer S, Kolarova V, Fraedrich E et al. 2016. Autonomous driving: The impact of vehicle automation on mobility behaviour. Institute for Mobility Research. Berlin, Germany.

168 Truong L T, De Gruyter C, Currie G et al. 2017. Estimating the trip generation impacts of autonomous vehicles on car travel in Victoria, Australia. Transportation, 44(6): 1279-1292.

169 U.S. Department of Transportation. 2015. Critical reasons for crashes investigated in the National Motor Vehicle Crash Causation Survey. National Highway Traffic Safety Adiministration Center for Statistics and Analysis. Washington DC, U.S.A.

170 Bertoncello M, Wee D. 2015. Ten ways autonomous driving could redefine the automotive world. McKinsey & Company.

171 Overly S. 2017. Uber suspends testing of self-driving cars after crash. The Sydney Morning Herald. Australia.

172 Greenberg A. 2015. After Jeep hack, Chrysler recalls 1.4M vehicles for bug fix. Wired.

173 Bogle A. 2018. Driverless cars and the 5 ethical questions on risk, safety and trust we still need to answer. Australian Broadcasting Corporation. Australia.

174 Bonnefon J-F, Shariff A, Rahwan I. 2016. The social dilemma of autonomous vehicles. Science, 352(6293): 1573.

175 Ackerman E. 2016. People want driverless cars with utilitarian ethics, unless they're a passenger. IEEE Spectrum.

176 Goodall N J. 'Machine ethics and automated vehicles' In: Road Vehicle Automation. Springer; 2014.

177 Gerdes J C, Thornton S M. 'Implementable ethics for autonomous vehicles' In: Maurer Markus *et al.* Autonomes Fahren: Technische, rechtliche und gesellschaftliche Aspekte. Berlin, Heidelberg: Springer Berlin Heidelberg; 2015.

178 Commission N T. Topics: Safety Assurance System for Automated Vehicles.

179 Commission N T. Current projects: Motor accident injury insurance and automated vehicles.

180 National Transport Commission. 2018. Safety assurance for automated driving systems: Decision regulation impact statement. National Transport Commission. Melbourne, Australia.

181 Reyes O C, Vera-Rodriguez R, Scully P et al. 2018. Analysis of spatio-temporal representations for robust footstep recognition with deep residual neural networks. IEEE.

182 QANTAS. 2018. Facial recognition. QANTAS, Travel Advice.

183 Department of Home Affairs. Why are we using face recognition technology in arrivals SmartGate? Australian Government.

184 Times of India. 2018. Delhi: Facial recognition system helps trace 3,000 missing children. Times of India. India, Dehli.

185 Dockrill P. Thousands of Vanished Children in India Have Been Identified by a New Technology. Sciencealert. 1 May 2018.

186 Eyers J. 2017. Westpac Testing AI to monitor staff and customers. Australian Financial Review.

187 National Health and Medical Research Council. 2007 (updated 2018). National Statement on Ethical Conduct in Human Research.

188 Big Brother Watch. 2018. Face Off: The Lawless Growth of Facial Recognition in UK Policing.

189 Davies B, Dawson A, Innes M. 2018. How facial recognition technology aids police.

190 Georgetown Law Center. 2016. The perpetual line-up: Unregulated police facial recognition in America. Georgetown Law Center on Privacy and Technology. U.S.A.

191 Frey C B, Osborne M A. 2013. The future of employment: How susceptible are jobs to computerisation. Technology Oxford Martin Programme on the Impacts of Future: Oxford.

192 Durrant-Whyte H, McCalman L, O'Callaghan S et al. 2015. The Impact of Computerisation and Automation on Future Employment. Committee for Economic Development (CEDA).

193 Nedelkoska L, Quintini G. 2018. Automation, skills use and training. Working Paper Number 202. Organisation for Economic Cooperation and Development Paris.

194 Nesta. 2017. The Future of Skills: Employment in 2030. Nesta (https://www.nesta.org.uk). London.

195 OECD. 2018. Transformative technologies and jobs of the future. Canadian G7 Innovation Ministers Meeting. The Organisation for Economic Cooperation and Development. Paris.

196 AlphaBeta. 2019. Future Skills - To adapt to the future of work, Australians will undertake a third more education and training and change what, when and how we learn. Prepared by AlphaBeta for Google Australia. Sydney.

197 Workplace Gender Equality Agency. 2018. Professional, Scientific and Technical Services summary.

198 Robinson C, McKaige B, Barber M et al. 2016. Report on the national Indigenous fire knowledge and fire management forum: Building protocols from practical experiences Darwin, Northern Territory 9th–10th February 2016 CSIRO and Northern Australia Environmental Resources Hub. Australia.

199 Janke T. 2018. Legal protection of Indigenous Knowledge in Australia.

200 Australian Institute of Aboriginal and Torres Strait Islander Studies. 2012. Guidelines for ethical research in Australian Indigenous studies. Australian Institute of Aboriginal and Torres Strait Islander Studies. Canberra, Australia.

201 Reed C. 2018. How should we regulate artificial intelligence? Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 376(2128): 1-12.

202 Government of Canada. Algorithmic impact assessment (v0.2). Government of Canada. Canada.

203 Office of the Australian Information Commissioner. 2014. Guide to undertaking privacy impact assessments. Australia.

204 Office of the Australian Information Commissioner. Privacy impact assessment eLearning.

205 Varshney K. 2018. Introducing AI Fairness 360. IBM.

206 Fairness Accountability and Transparency in Machine Learning. Principles for accountable algorithms and a social impact statement for algorithms. Fair and Transparent Machine Learning.

207 The National Health and Medical Research Council, The Australian Research Council and the Australian, Vice-Chancellors' Committee. 2007 (Updated May 2015). National statement on ethical conduct in human research. National Health and Medical Research Council. Canberra.

208 Fair Work Ombudsman. Best practice guides. Australian Government. Australia.

209 Erdelyi O, Goldsmith J. 2018. Regulating artificial intelligence: A proposal for a global solution. Conference on Artificial Intelligence, Ethics and Society.

210 Administrative Review Council. 2004. Automated assistance in administrative decision making: Report no. 46. Administrative Review Council. Canberra.

211 Australian Medical Association. 2016. AMA code of ethics. Australian Medical Association. Australia.

212 Department of Education and Training. Licensing: Process for gaining a current, identified Australian occupational licence. Australian Government. Australia.

213 Department of Mines Industry Regulation and Safety. Motor vehicle repairer's certificate. Government of Western Australia. Australia.

214 Engineers Australia. Program accreditation overview. Available from: https://www.engineersaustralia.org.au/About-Us/Accreditation.

215 Data Governance Australia. Leading Practice Data Principles. Available from: http://datagovernanceaus.com.au/leading-practice-data-principles/.

216 Department of Industry Innovation and Science. 2015. National innovation and science agenda. Australian Government. Australia.

217 IP Australia, Department of Industry Innovation and Science. 2018. The newly updated IP Toolkit is now live. IP Australia. Australia.

218 Australian Technology Network of Universities. 2017. Enhancing the value of PhDs to Australian industry. Australian Technology Network of Universities. Australia.

219 Gasser U, Budish R, Ashar A. 2018. Module on setting the stage for AI governance: Interfaces, infrastructures, and institutions for policymakers and regulators. International Telecommunications Union.

Appendix Stakeholder and Expert Consultation

Targeted consultations with 91 invited representatives from universities and institutes, industry and government were conducted across four Australian capital cities (Melbourne, Brisbane, Perth and Sydney). The workshops were a key component to developing a cohesive and representative narrative that accurately captured the perspectives and priorities of various Australian stakeholders. In addition to the four consultation sessions, advisory and technical expert groups were engaged in the development of this report (Figure 7).

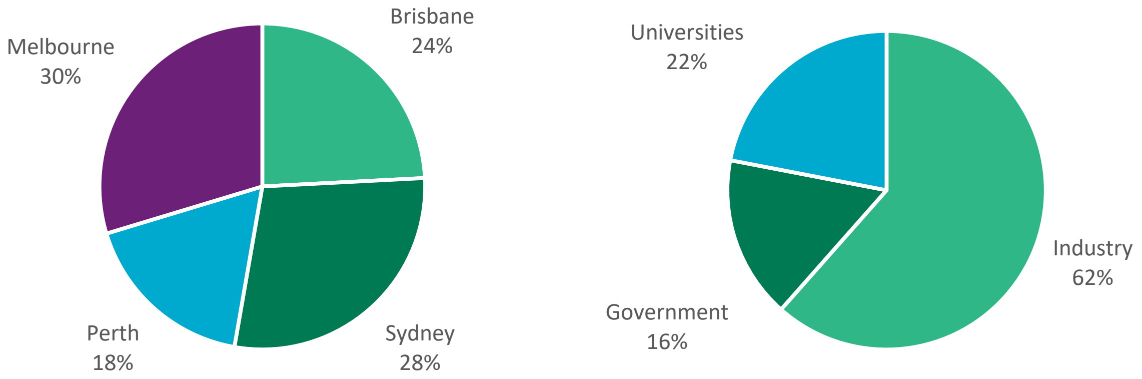


Figure 7. Pie charts of consultation attendee demographics

Note: A total of 91 persons were consulted at the workshops. Additional consultations were held with other industry, research and government experts.

Consultation approach

The consultations were run as informal workshops focused on a collaborative and generative approach. Participants were encouraged to both share their ideas, and develop and collaborate with others.

The workshops began with a presentation introducing the topic of AI as well as the objectives and structures of the AI Ethics Framework. Following the presentation participants were given opportunity to question and interrogate the approach of both reports, and where appropriate, that feedback has been integrated into the reports.

Following the presentation, in the first discussion session, participants were given the opportunity to group together and discuss their perspectives on the biggest opportunities for Australia in the use and adoption of AI and the factors that could enable or inhibit that adoption. In the second discussion session participants were asked to consider their perspectives on risk mitigation and measures needed to ensure wide-spread societal benefits from the adoption of AI. Each of the workshops provided robust dialogue and diverse perspectives across both discussion sections that resulted in an informative snapshot of Australia's unique AI opportunities and challenges.

Key themes

- Prioritization is key: Participants acknowledged the need to prioritise investment in AI on focussed, strategic areas to take advantage of Australia's unique opportunities and address its challenges.

  - Whilst the opinions on which domains should be of focus varied, it was reaffirmed that this investment should be conducted in conjunction with other major initiatives across the Australian landscape.
  - There were several discussions on the potential for Australia to be a leader in AI integration, in addition to AI development, particularly across primary the industries.
  - Suggestions were also made about Australia's potential to play a world leading role in ethical and responsible AI.
- Need for skilled workers: All participants recognised a significant knowledge gap in the current workforce as a whole, and that future-proofing skills and curriculum for the next generation of knowledge workers needed to be addressed.
  - The responsibility for ensuring Australia has a strong technically-enabled workforce was seen to be a shared responsibility across all sectors.
- Collaboration and multidisciplinary approaches are required: Across all sectors, in each city, the need for Australia to develop much stronger collaboration within and between sectors was stated and reiterated.
  - There is strong appetite to connect between the industry and researchers involved in the workshops, but a lack of infrastructure to encourage this engagement and remove friction
  - The need for initiatives that could help coordinate how sectors could work together on projects, rather than compete for them.
  - The multi-disciplinary nature of AI was discussed along with the need for collaborative approaches to ensure that Australia can optimise their use and adoption of AI in the most positive way.
- Data governance: Discussion was focussed on the steps that need to be made to ensure that data privacy regulations are adhered to, without limiting development or adoption of AI.
  - In addition the need to address the lack of large datasets in Australia and how this will affect our ability to compete in AI development on a global scale
- Embracing AI: There was a healthy appreciation of strategies to mitigate risks, and demystify artificial intelligence. These included ensuring that AI adhered to ethical principles such as fairness and transparency.
  - It was raised that over-regulation could limit the opportunity for Australia to play a leading role, and having nimble and responsive frameworks would serve better.
  - There was discussion around the need for a cohesive nationwide approach to addressing ethical issues associated with AI use and adoption.
  - In every session the need to use AI for good was suggested and discussed by participants. In particular the need to use AI to address Australia's sustainability and environmental issues was discussed in each session.

CONTACT US

t 1300 363 400
+61 3 9545 2176

e csiroenquiries@csiro.au

w www.data61.csiro.au

WE DO THE EXTRAORDINARY EVERY DAY

We innovate for tomorrow and help improve today – for our customers, all Australians and the world.

WE IMAGINE
WE COLLABORATE
WE INNOVATE

FOR FURTHER INFORMATION

Stefan Hajkowicz

t +61 7 3833 5540

e stefan.hajkowicz@data61.csiro.au

w www.data61.csiro.au

Appendix 3: Ethical Artificial Intelligence in the Australian Signals Directorate

Australian Government
Australian Signals Directorate

ASD

Ethical Artificial Intelligence in the Australian Signals Directorate

January 2023

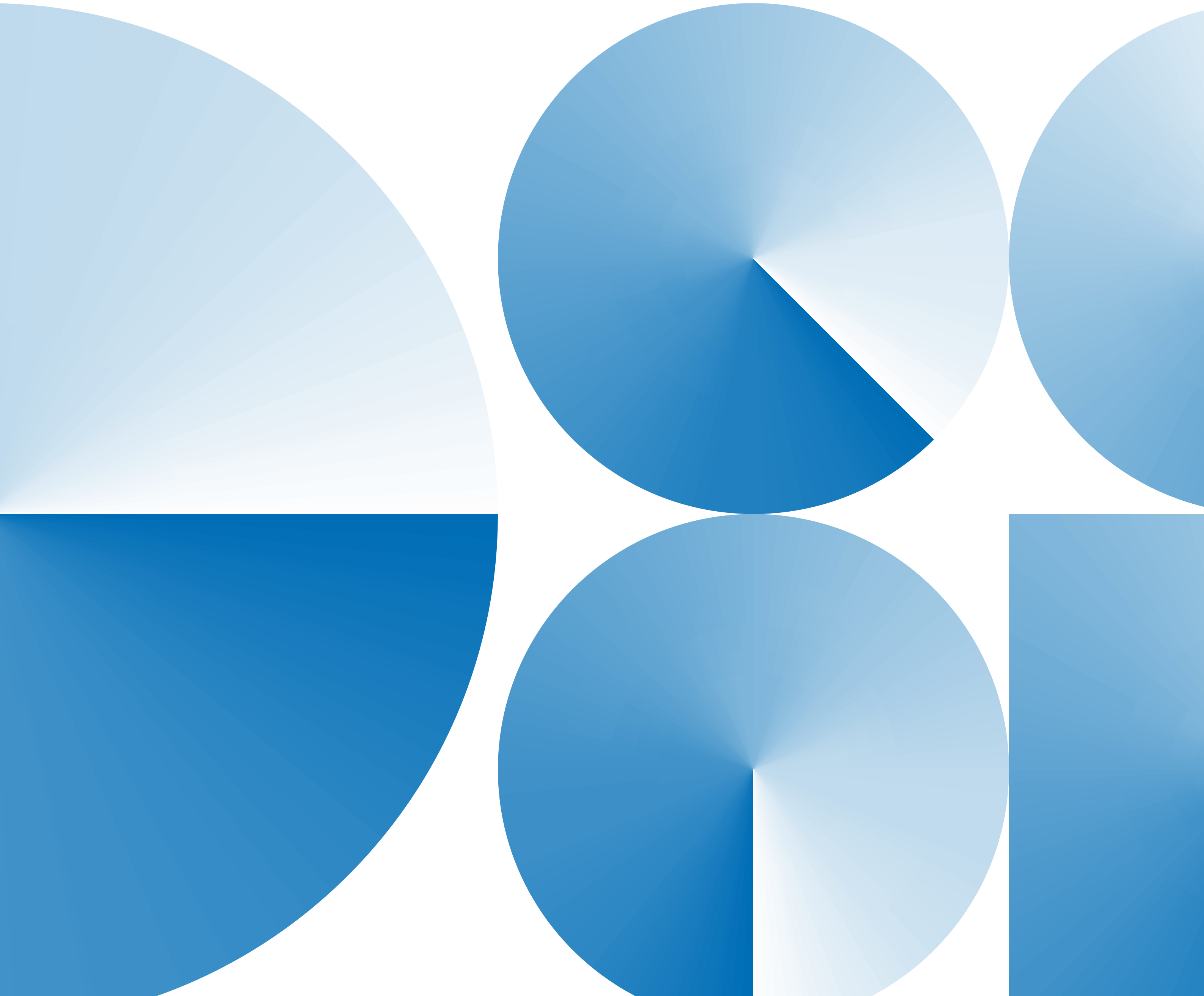


Table of Contents

Foreword

The environment in which the Australian Signals Directorate (ASD) operates is becoming increasingly challenging. Today, Australia's region – the Indo-Pacific – is far more complex and far less predictable than at any time since the Second World War. Mastery of innovation and technology will play an important role in helping avoid miscalculation and conflict.

I get asked a lot about how ASD will contend with the technological challenges of the future. My answer is simple – with the same innovation, imagination and creativity that we have used to deal with challenges for the last 75 years. We have a proud history of adapting at the pace that technology evolves, and for the global adoption of artificial intelligence (AI), our approach will be no different.

Harnessing the full potential of AI will deliver strategic advantage for Australia, and will enable ASD analysts to focus on the greatest threats. While ASD has always provided intelligence that delivers advantage to the Australian Defence Force and protected the security of its communications, AI will enable ASD to conduct rapid intelligence, helping provide our people with the best possible tools to complete their mission.

By necessity, much of ASD's work happens in secret, but core to our values is ensuring that we not only meet our legal obligations, but operate within community expectations of propriety and ethics. The AI Ethical Framework outlined in this paper demonstrates how we apply AI in a lawful, secure, accurate, fair and accountable manner. Human judgement will remain at the centre of our decision-making, and our talented people will always be the critical component in our efforts to protect Australia's interests.

Rachel Noble PSM

Director-General ASD

Introduction

ASD and its mission-focused culture was born out of support to the Allied cryptologic effort during the Second World War. This global enterprise included some of the world's finest mathematicians and cryptanalysts who developed the foundations of digital computing and programming. Staff from ASD's forebears contributed to the Allied war effort led by Alan Turing OBE and other great minds to break German and Japanese codes in the 1930s and 1940s. This allowed the Allies to read enemy communications and gain crucial advantage during the war. A leader in computer science and artificial intelligence (AI), Turing went on to develop the 'Turing Test', a method for assessing whether machines can think[1].

For more than 75 years, ASD has collected intelligence about foreign adversaries while keeping Australia's national secrets safe. From decrypting Japanese radio signals during the Second World War to our role today at the forefront of contemporary signals intelligence and cyber security, ASD has always leveraged state-of-the-art technologies. The use of AI represents the next phase in maintaining and securing a competitive advantage.

Advances in technology are changing the way Australians live, work and communicate. They create new jobs, drive economic growth and boost competiveness. But with technological advancement come different threats and adversaries who seek to undermine Australia's interests.

AI is a form of software that can learn to solve problems at a scale and speed impossible for humans, and it is all around us. It guides our internet searches, helps point us in the right direction when using satellite navigation, and enables technologies as diverse as voice transcription and medical screening for cancers. AI is used widely by businesses and industry to detect and recommend preventative maintenance, shape complex logistics networks, and target advertising and content on social media.

At ASD, AI has a unique role to play in supporting our highly skilled people. It can automate tasks that machines excel at, which enables our people to focus on the things that only people can do. The interaction between machine automation and human decision-making will enable us to better develop and tailor solutions to emerging security challenges.

As we continue to develop our AI capability, ASD will strive to guard against the inherent risks as they relate to the collection, analysis and assessment of intelligence as well as the management of cyber threats.

Implementing AI in a national security context represents particular challenges. By necessity, much of ASD's work is done in secret, but we still adhere to the social and ethical standards expected of us by the Australian community. To ensure this, ASD will adapt and grow its governance of AI as community expectations and standards evolve. In this context, establishing a principles-based approach to AI ethics ensures ASD presents not only an ethics governance framework but also a pathway for practical implementation.

[1] Oppy, G., and Dowe, D., 'The Turing Test' in *The Stanford Encyclopaedia of Philosophy* (Winter 2021 edition), Edward Zalta (editor), https://plato.stanford.edu/archives/win2021/entries/turing-test/ (retrieved 02 November 2022).

ASD’s Purpose

ASD defends Australia from global threats and helps advance Australia’s national interests. We do this by mastering technology to inform, protect and disrupt.

- Inform: by the covert acquisition of foreign information not publicly available (signals intelligence).
- Protect: by comprehensively understanding the cyber threat, providing proactive advice and assistance to improve the management of cyber risk by governments, businesses and the community.
- Disrupt: by applying our offensive cyber capabilities offshore to support military operations and disrupt cybercrime.

The Australian Government continues to make significant investments in ASD’s foreign signals intelligence, cyber security role and offensive cyber capabilities. This includes an AUD $9.9 billion investment in cyber and intelligence capabilities over the next decade through the REDSPICE (Resilience – Effects – Defence – SPace – Intelligence – Cyber – Enablers) initiative. REDSPICE responds to the deteriorating strategic circumstances in our region, which are characterised by rapid military expansion, growing coercive behaviour and increased cyber attacks. Through REDSPICE, ASD will deliver forward-leaning capabilities essential to maintaining Australia’s strategic advantage and capability edge over the coming decade and beyond. This includes advanced AI, machine learning and cloud technology in support of ASD’s mission.

What is AI?

AI is a collection of technologies that can be employed to perform tasks that would typically be done by a human, including at a scale, speed and complexity that humans cannot do. AI can also enable and augment human teams to:

- support better decision-making
- discover new ways of operating
- increase efficiency and resilience.

AI may also refer to process automation or mathematical algorithms. It may also include machine learning, which is the ability for a computer to ingest and learn from data in order to make predictions or identify patterns. An AI solution will often incorporate all of these technologies.

AI is essential in assisting ASD to meet the challenges and changes in Australia's strategic environment. These include:

- dealing with the increasing complexity and volume of data
- responding to new developments in technology
- providing ASD with the best possible tools to enable their mission
- adapting to changes in the environment in a way that rules-based approaches cannot.

AI technologies evolve rapidly, which is why ASD recognises the importance of being agile in developing AI solutions, and responsive in adapting its processes, guidelines and concepts.

AI at ASD

ASD already leverages AI to support a variety of its functions, in particular cyber security, intelligence analysis and support to the Australian Defence Force (ADF).

Cyber security

Cyber security and reinforcing online resilience is a national priority. Australia's prosperity makes us attractive to cyber criminals and malicious cyber activity is growing. Rapid exploitation of critical public vulnerabilities has become the norm, and critical infrastructure networks are increasingly being targeted worldwide. AI can help ASD to protect government networks, defend critical infrastructure, and advise all Australian internet users by:

- providing early detection of a cyber incident, whether the incident was successful or not
- rapidly identifying damage and/or loss of capability to ensure application of appropriate mitigations
- suggesting and prioritising disruption and response options, particularly for automated systems
- minimising disruption to operations and establishing resilience within networks to ensure limited loss of function during and after a cyber incident.

Intelligence analysis

Intelligence analysis is about finding the needle in the haystack to protect Australia and our national interests. Over the last two decades, the total volume of information created, captured, copied, and consumed globally has grown exponentially. So while the needle has not changed size, the haystack has grown tremendously. AI will be central to ASD in rising to meet this challenge by:

- filtering substantial amounts of data to identify high interest communications, allowing analysts to concentrate on Australia's greatest threats
- enabling analysis that would otherwise be too difficult or voluminous
- supporting decision-makers with rapid situational awareness, verification of results and increased confidence
- providing better resource management for the collection, storage and analysis of data.

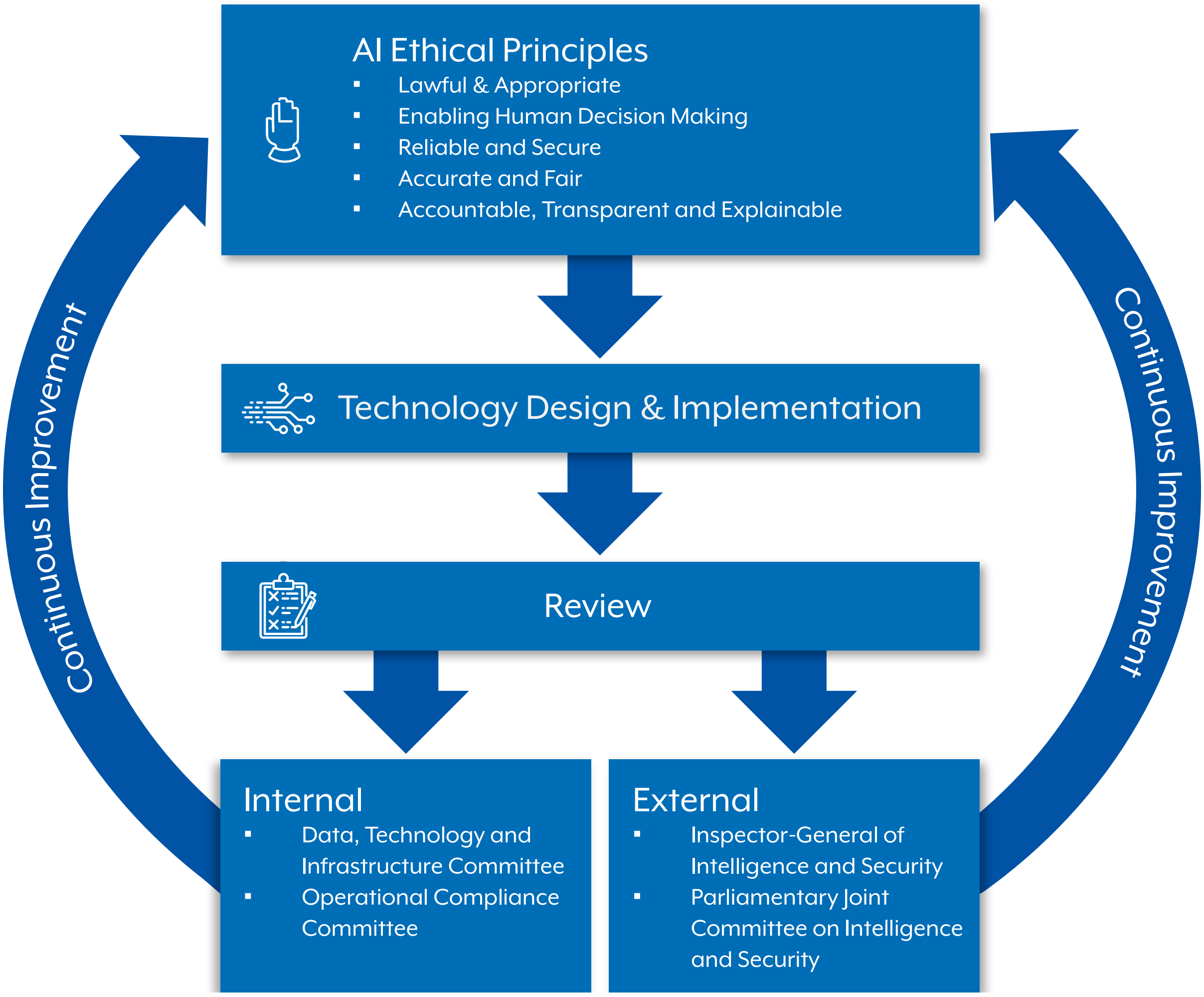
Support to the ADF

ASD has a long history of supporting Australian military operations. Today, ASD supports ADF operations around the globe to enable warfighting and protect our people and assets. ASD draws on its deep technical expertise to help the ADF stay ahead of technological advancements in the region. AI will support this mission by:

- providing ADF platform protection through the rapid analysis of threat data
- fusing and analysing multi-source data to enable faster and more robust decision-making.

ASD's Ethical AI Framework

The foundation of ASD's Ethical AI Framework (Figure 1) is the AI Ethical Principles. The principles will drive future AI technology design and implementation in ASD. ASD will leverage existing processes for internal and external review to drive continuous improvement.



AI Ethical Principles

- Lawful & Appropriate
- Enabling Human Decision Making
- Reliable and Secure
- Accurate and Fair
- Accountable, Transparent and Explainable

Technology Design & Implementation

Review

Internal

- Data, Technology and Infrastructure Committee
- Operational Compliance Committee

External

- Inspector-General of Intelligence and Security
- Parliamentary Joint Committee on Intelligence and Security

FIGURE 1: ASD Ethical AI Framework

AI Ethical Principles

We recognise that the promise of AI raises questions around how ASD manages complex systems that can have a direct impact on the privacy and security of Australians. The AI Ethical Principles below help answer these questions. They also provide appropriate guidance for ASD staff involved in the development, deployment and maintenance of AI systems.

The principles

- **Lawful and appropriate**: ASD's development, management, and deployment of AI capabilities are consistent with the legislation, policies, processes and frameworks that govern ASD's functions and protect the privacy of Australian citizens.
- **Enabling human decision-making**: Human assessment and judgement will remain central. ASD will use AI to support and enable its highly-skilled workforce to best fulfil ASD's functions in protecting Australia and Australians from threats.
- **Reliable and secure**: AI capabilities developed by ASD will be managed to ensure that they are reliable, continue to meet their intended purpose, and remain protected from external interference.
- **Accurate and fair**: ASD will endeavour to remove unintended bias from its AI systems so that they produce results that are balanced, accurate and fair.
- **Accountable, transparent and explainable:** AI-based capabilities will have human oversight and control, with clear accountabilities in place for all stages of the development life-cycle. ASD is committed to meeting the ethical need for its decisions and actions to be accountable, transparent and explainable to ensure it operates appropriately and proportionately. This will be balanced with the need to protect sensitive equities to ensure ASD is able to continue to perform its critical intelligence and security functions.

The AI Ethical Principles are an important step in a program of work within ASD to build world-class AI capabilities that support our organisation to defend Australia from global threats and advance Australia's national interests. ASD will continue to develop the technical, policy and governance frameworks required to help us manage our AI capabilities in line with these principles and we will continue to invest in our most valuable asset, our workforce, to fully realise the benefits that AI offers us as an organisation.

Implementing AI ethically

There are challenges associated with the fair and transparent use of AI in any industry. Meeting these challenges in a national security context is particularly demanding because ASD cannot always be completely transparent about its operations. Our AI Ethical Principles are the first step in addressing these challenges. Some specific challenges that ASD seeks to address, through the application of these principles, are:

- **Minimising unintended bias**: AI models can reflect biases made by system creators in their training data and design decisions. This can lead to biased results such as disadvantaging certain demographics. Appropriate guidelines and governance arrangements will be developed to ensure that ASD minimises unintended bias and that the true and intended outcome of an AI system aligns with ASD's licence to operate.
- **Appropriate use of AI models**: Using a model for a purpose for which it was not originally designed can have unintended consequences that impact on individuals. Even when a model is used for its intended purpose, model drift over time may introduce biases or inaccuracy, meaning that it is no longer suitable for its intended purpose. Governance related to the ongoing applicability of AI models and standards for accuracy and reliability will be developed to ensure the continued appropriate use of AI systems in ASD.
- **Enabling human decision-making**: Effective human-machine teaming is important for users to understand how to make informed decisions based on AI system outputs, and to maintain trust in AI systems. ASD's workforce will have a solid foundational knowledge of how AI works, enabling them to use the right model on the right data and the confidence to judge the level of certainty of the results.
- **Securing AI systems**: Attacks by adversaries could affect the behaviour of AI systems and the decisions made as a result of their outputs. Standards for system monitoring, reliability and life-cycle management will be developed to safeguard the health and security of AI systems.

Governance and oversight of AI

As an intelligence agency, ASD has been entrusted with sensitive powers and must be accountable to the public, through the Government, for everything it does. This accountability is exercised through multiple external oversight bodies. In particular:

- the Inspector-General of Intelligence and Security (IGIS), who has the powers of a standing Royal Commission – provides independent assurance that ASD acts legally, with propriety and consistent with human rights
- the Parliamentary Joint Committee on Intelligence and Security (PJCIS), which provides oversight of ASD's administration, expenditure and enabling legislation.

Additionally:

- the Independent National Security Legislation Monitor (INSLM) independently reviews the operation, effectiveness and implications of national security and counter-terrorism laws in which ASD has equities
- the Auditor-General and the Australian National Audit Office (ANAO) provides assurance that ASD is operating and accounting for its performance in accordance with the Parliament's intent through independent audit reporting.

Internally, the Data, Technology and Infrastructure Committee (DTIC) is ASD's senior data and technology decision-making committee. The DTIC ensures that ASD makes evidence-based decisions on data management and meets the highest standards of portfolio management practice and technology delivery.

Additionally, ASD's Operational Compliance Committee helps ensure ASD remains compliant in its intelligence activities, cyber security activities and offensive cyber operations, as well as ensuring ASD establishes and maintains best practice in its operational policies and procedures. This includes compliance with ASD's Ethical AI Framework and AI Ethical Principles.

ASD has also implemented recommendations from the Comprehensive Review of the Legal Framework of the National Intelligence Community (Comprehensive Review) to strengthen governance and oversight of AI. This includes providing the PJCIS with an annual submission on the development of its AI-based intelligence capabilities and maintaining human involvement in significant or difficult AI-based decisions.

ASD will continue to refine and adapt its enterprise processes and governance structures as AI technologies evolve. Initiatives, such as a catalogued and discoverable database of AI tools and models that have been proven to meet our AI Ethical Principles, will ensure that these technologies are not only harnessed across all of ASD's missions, but are also implemented in a consistent, legal and ethical manner. Such initiatives may also enhance the sharing of AI technologies and management practices with ASD's Five-Eyes Partners.

Partnerships

ASD has a long and proud tradition of adopting new technologies and using them to drive better intelligence and security outcomes for Australia. We have operated in the slim area between the difficult and the impossible since our inception, and the Five-Eyes Partnership, forged during the Second World War, has brought immeasurable benefit, including in the field of AI and machine learning. Without this partnership, ASD would never have been able to independently achieve for Australia the advances we have in shared technology, innovation, capability and 'reach'; we are stronger together.

Through continued collaboration with our Five-Eyes Partners, and with industry and academic partners, ASD's AI capabilities can advance in step with emerging technologies. This will drive improvements in quality, efficiency, and timeliness across all aspects of our business including the triage of large volumes of data to identify high value intelligence, the automation of routine tasks and the early detection of anomalous cyber activity. ASD's approach is to build a team that operates seamlessly with these new technologies. It is the interaction between subject matter experts and AI capabilities that will unlock the true potential of partnership.

In this spirit, the ASD-ANU Co-Lab partnership brings together problem-solvers from a variety of disciplines to conduct complex research on Australia's toughest national security problems. The CoLab brings the next generation of talented scientists and mathematicians into ASD's partnerships, and together we are tackling projects on topics such as machine learning and AI ethics, cryptography, cyber security and vulnerability analysis.

ASD will further extend its partnership with academia through the establishment of an AI Hub in 2023 as part of REDSPICE. The AI Hub will focus on acquiring and integrating AI technologies into production systems, and will be a driving force in partnering industry with academia to bring cutting-edge research to new product development.

Conclusion

With the exponential growth in information globally, ASD must adapt and innovate to meet its strategic objectives. Harnessing the full potential of AI will deliver strategic advantage for Australia, freeing up ASD's analysts to focus on the greatest threats. ASD's international, industry and academic partnerships will help drive the development and implementation of cutting-edge AI applications. ASD will apply these capabilities in a way that is secure, trusted and ethical, and it will be as transparent as possible about their use.

asd.gov.au

As Australia’s national science agency and innovation catalyst, CSIRO is solving the greatest challenges through innovative science and technology.

CSIRO. Unlocking a better future for everyone.

www.csiro.au