ARTIFICIAL INTELLIGENCE AND RECORDKEEPING

Research Paper

Abstract

This paper makes recommendations for a Victorian government approach to managing records created by or using artificial intelligence (AI) technologies. The paper includes possible resources and methods for implementation within a Victorian jurisdiction and current methods of regulating A1 from across the globe. Also included are common types of AI in use, their known risks and harms in relation to records, and a basic glossary.

Xander Hunter (PROV)
xander.hunter@prov.vic.gov.au

Public Record Office Victoria

VICTORIA State Government

Table of Contents

Executive Summary

This research paper makes recommendations for a Victorian government approach to managing records created by or using artificial intelligence (AI) technologies. Recommendations are used to inform Public Record Office Victoria (PROV) advice to public offices on capturing and managing accurate and trustworthy records created by, or through use of AI technologies.

It explores:

- two major international approaches for regulating AI technologies. They are:
    - the *European Union Artificial Intelligence Act* and associated products, and
    - the *United States Blueprint for an AI Bill of Rights*, and associated *National Institute of Science and Technology AI Risk Management Framework* and related products.
- approaches that have so far emerged within Australasia in relation to assessing and documenting AI technologies.
- existing tools for assessing and documenting AI technologies that could be used by Victorian government bodies as external resources.
- guidance already available from PROV regarding AI technologies in relation to documenting approval processes.

PROV acknowledges that:

- a Victorian AI Hub / Centre of Excellence (CoE) is in development, but is not yet in operation.
- a Victorian Government Generative AI Policy is in development by the Data Branch of the Department of Government Services.

The recommended approach for creating and managing accurate and trustworthy records created by or through the use of AI technologies across Victorian government is as follows:

- that general advice be developed by PROV in the form of a topic page that covers:
    - what AI technologies are
    - what needs to be considered regarding creating, capturing, and managing records
    - how it links with existing PROV advice, including the current *Approval Processes Policy*[1].
- that new policy advice regarding Generative AI technologies (including Large Language Models or LLMs) be produced to clarify what is required from a recordkeeping perspective with a focus on accuracy.
    - The existing *Approval Processes Policy* covers automotive and machine learning AI technologies that make and / or action decisions, and therefore focuses on ensuring transparency.
    - Generative AI technologies produce convincing and human-like content, such as minutes of meetings, reports, and artwork, based on what the user is looking for and have been known to invent content. The Generative AI Technologies guidance would therefore relate to the creation and capture of full and accurate records of business.

[1] https://prov.vic.gov.au/recordkeeping-government/document-library/approvalprocessespolicy-approval-processes-policy

- that existing assessment frameworks and associated tools for documenting AI technologies across the globe be connected to the general AI guidance as useful external resources for Victorian government agencies.
- that the PROV advice be reviewed and updated once an approach from the AI Hub / CoE has become available.

Introduction

"Artificial intelligence ('AI') is a collective term for machine-based or digital systems that use machine or human-provided inputs to perform advanced tasks for a human-defined objective, such as producing predictions, advice, inferences, decisions, or generating content."[2]

Purpose and Scope

The purpose of this research paper is to broadly outline approaches for the trustworthy management of records created by or through the use of artificial intelligence (AI) technologies in order to recommend an approach for Victorian government recordkeeping.

It explores:

- two major international approaches for regulating AI technologies. They are:
    - the *European Union Artificial Intelligence Act*[3] and associated products, and
    - the *United States Blueprint for an AI Bill of Rights*[4], and associated *National Institute of Science and Technology AI Risk Management Framework* [5]and related products.
- approaches that have so far emerged within Australasia in relation to assessing and documenting AI technologies.
- existing tools for assessing and documenting AI technologies that could be used by Victorian government bodies as external resources.
- guidance already available from Public Record Office Victoria (PROV) regarding AI technologies in relation to documenting approval processes (which includes decision-making AI technologies).

Recommendations are used to inform PROV advice to public offices on capturing and managing trustworthy and accurate records about, created by, or through use of AI technologies.

Context

AI technologies can contribute to the process of making, managing, and using records, can be something that creates content, can make decisions, and can take actions on records. There are different kinds of AI technologies that come with different sets of challenges (see *Appendix One: Common Types of AI*). AI technologies are progressing quickly, and the space is constantly evolving.

A multi-disciplined and collaborative approach to managing records of AI is necessary for full and accurate records to be created, kept, and appropriately disposed of. Ideally, this would be in line with a jurisdiction wide and general set of AI technologies regulations. In the absence of Victorian regulation, PROV advice in relation to recordkeeping and AI technologies would clarify any recordkeeping actions required to specifically address accuracy and transparency in relation to use of AI technologies. It would also help to guide discussion between records managers and related professional areas, such as data management and information technology.

The paper acknowledges that a Victorian AI Hub / Centre of Excellence is in development but is not currently operational.

---

[2] Solomon, L., & Davis, N., (2023) *The State of AI Governance in Australia*, Human Technology Institute, The University of Technology Sydney

[3] https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.html

[4] https://www.whitehouse.gov/wp-content/uploads/2022/10/Blueprint-for-an-AI-Bill-of-Rights.pdf

[5] https://www.nist.gov/itl/ai-risk-management-framework

AI Technologies and Recordkeeping

Businesses are encouraged to improve efficiency and timeliness of operations, and AI technologies provide tools that can help with this. Ethical principles have emerged in conjunction with AI technologies to help identify and mitigate potential harms caused.[6] Stakeholders are increasingly aware of the impact that use of AI technologies have on the decisions and actions that directly affect them. [7] Using AI technologies without thought as to their possible impact on all stakeholders involved has been demonstrated to cause harm (see *Appendix Two: Risk and Harms and their Mitigation*). Records provide documented evidence of what transpired.

Main types of AI

Different types of AI have different purposes. For example:

- Automation AI technologies undertake **actions** in accordance with specific parameters and data sets available.
- Machine learning and other decision-making AI technologies **make decisions** in accordance with specific parameters and data sets available.
- Generative AI including Large Language Models (LLMs) **create new content** in accordance with specific parameters and data sets available.
- Combination AI technologies can do **all of the above**: create content, form decisions on that content, and then perform actions on that content in accordance with specific parameters and data sets available.

Recordkeeping Considerations

When considering requests for evidence, businesses need to address bias, accuracy, and transparency (see *Appendix Three: Good Recordkeeping and AI*, *Appendix Four: Ethical Principles and Recordkeeping* and *Appendix Five: Glossary*). This includes:

- What AI technologies were used.
- How they work.
- What assessments to identify possible harm were used, when they were used, the results of the assessment, and what was done to mitigate possible harms.
- How to document the technologies in order to best address transparency, accuracy, and bias.

Assessing AI technologies for risk and potential harms is not a simple process. There is no standardised way of assessing for risk mitigation, no standardised vocabulary, different ethical principles depending on the jurisdiction, different stages of AI technology lifecycle to consider, and multiple parties involved.[8] An ecosystem of risk assessment covering different sectors and domains,

---

[6] For example: 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. And Digital New South Wales, NSW Government (2022, March). *NSW Artificial Intelligence Ethics Policy*. Digital New South Wales, NSW Government. Accessed August 2023 https://www.digital.nsw.gov.au/policy/artificial-intelligence/artificial-intelligence-ethics-policy

[7] https://www.ajl.org/

[8] Brennan, Jenny (2023), AI Assurance? Assessing and Mitigating Risks Across the AI Lifecycle. Ada Lovelace Institute, London, England. Accessed September 2023: https://www.adalovelaceinstitute.org/report/risks-ai-systems/.

and across the various stages of the lifecycle, is recommended and should include recordkeeping in accordance with accepted standards. The development of recordkeeping-specific advice from PROV would address this gap, and could be tailored to fit within any Victorian-wide regulatory framework.

Regulatory Frameworks and AI

Regulatory frameworks regarding AI technologies can largely be divided between two camps:

- Specific AI legislation, regulation, and associated frameworks.
- Technology neutral legislation and regulation that encourages inclusion of AI technologies within existing frameworks.

Currently all frameworks are based on some form of risk assessment, whether undertaken by external bodies or through self-assessment. They are also tied to an ethical framework in relation to how AI technologies are to be designed, implemented, and used. The ethical frameworks are aligned around a concept of trustworthiness. Frameworks seek to address business innovation and personal protection, with different frameworks tending to focus on one more than the other.

The European Union is a good example of implementing specific AI legislation and regulations. The USA is a good example of implementing technology neutral legislation and self-assessment frameworks. Both are explored in more detail, below.

European Union

The approach used by the European Union is to use legislation (the EU AI Act[9]) to regulate the developers of AI technologies, software companies that include AI technologies, companies that use AI software, and the public that use AI and are impacted by AI technologies. It is based around the use of standards and regulatory bodies to implement the requirements of the AI Act. It includes data requirements as well as recordkeeping requirements. The EU AI Act's concept of trustworthiness has been informed by the European Commission's *Ethics Guidelines for Trustworthy AI*.[10]

The EU AI Act[11] sets different rules for different risk levels regarding the development of AI, inclusion of AI in software or systems, and use of AI technologies by organisations and the public:

- Unacceptable risk results in the AI technologies' being banned, and include cognitive behavioural manipulation, social scoring, and biometric identification systems.
- High risk systems are AI systems that negatively affect safety or fundamental rights. All high-risk AI systems are to be assessed before being put on the market as well as throughout their lifecycle. High risk AI systems are to be divided into two categories (as specified below).
    - AI systems that are used in products falling under the EU's product safety legislation (eg toys, aviation, cars, medical devices and lifts).
    - AI systems falling into eight specific areas that will have to be registered in an EU database.
        - Biometric identification and categorisation of natural persons.
        - Management and operation of critical infrastructure.
        - Education and vocational training.

[9] https://artificialintelligenceact.eu/
[10] https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai
[11] summary based on European Parliament Website summary, accessed 16/08/2023: https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence

    - Employment, worker management and access to self-employment.
    - Access to and enjoyment of essential private services and public services and benefits.
    - Law enforcement.
    - Migration, asylum, and border control management.
    - Assistance in legal interpretation and application of the law.
- Generative AI technologies are to comply with the following transparency measures:
    - Disclose that the content was generated by AI.
    - Design the model to prevent it from generating illegal content.
    - Publish summaries of copyrighted data used for training.
- Limited risk AI systems should comply with minimal transparency requirements that would allow users to make informed decisions.
    - This means that after interacting with the applications, the user can decide whether they want to continue using it.
    - Users should be made aware when they are interacting with AI. This includes AI systems that generate or manipulate image, audio, or video content (for example deepfakes).

The AI Act has passed the EU parliament and is in the process of being implemented.

A conformity assessment framework (capAI) for the EU AI Act has been developed by researchers from the University of Oxford.[12] The tool is currently undergoing a validation process and is to be a living document that will be continually updated. CapAI is described as a procedure that provides practical guidance for organisations on how to translate the high-level ethics principles from the EU AI Act into criteria to help shape the design, development, deployment, and use of ethical AI. Its purpose is to demonstrate that the development and operation of AI systems are trustworthy.

A measure as to the transparency of some AI software is beginning to be, can it be implemented across Europe? For example, ChatGPT is under investigation or banned in multiple European countries as it is in breach of data protection laws. [13] As the EU AI Act enforces additional layers of regulation, AI technologies will come under increasing scrutiny.

Spain is the first country to implement the new EU AI Act within its legislator and regulatory frameworks. Many are watching to see what transpires, what challenges they come across and how successfully they address them.

USA

The US approach is described by the *Blueprint for an AI Bill of Rights*[14], which provides a national values statement and toolkit that is sector agnostic. It is to be used to inform building protections as

[12] https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4064091
[13] https://www.clydeco.com/en/insights/2023/07/chatgpt-faces-further-investigation-from-the-eu-an
[14] https://www.whitehouse.gov/wp-content/uploads/2022/10/Blueprint-for-an-AI-Bill-of-Rights.pdf

developed by experts across the globe into policy, practice, and technological design. The Blueprint is to be used when existing law or policy do not already contain guidance on AI.

The *Blueprint for an AI Bill of Rights* is made up of the following components:

- five value principles
    - *Safe and effective systems*: You should be protected from unsafe or ineffective systems.
    - *Algorithmic discrimination protections*: You should not face discrimination by algorithms and systems should be used and designed in an equitable way.
    - *Data privacy*: You should be protected from abusive data practices via built-in protections, and you should have agency over how data about you is used.
    - *Notice and Explanation*: You should know that an automated system is being used and understand how and why it contributes to outcomes that impact you.
    - *Human alternatives, consideration, and fallback*: You should be able to opt out, where appropriate, and have access to a person who can quickly consider and remedy problems you encounter.
- Notes on *Applying the Blueprint for an AI Bill of Rights*
- A *Technical Companion* that gives concrete steps that can be taken by many kinds of organisations to uphold the value principles. It explores each principle in three ways:
    - Why the principle is important, consisting of a brief summary of the problems that the principle seeks to address and protect against, including illustrative examples.
    - An outline of the expectations for design, development, implementation, and reporting.
    - Real-life examples of how these guiding principles can become reality, through laws, policies, and practices. It describes practical technical and sociotechnical approaches to protecting rights, opportunities, and access.

The Blueprint is perceived of as being a precursor to building standards, policies and whatever else is needed for ethical and responsible AI. It is intended to be flexible in its application so it can be implemented across the various States of the US, each of which has its own unique set of legislation. The US National Institute of Standards and Technology (NIST)has produced an *AI Risk Management Framework* (AI RMF), and associated Playbook that is aligned with the Blueprint.[15] The playbook links the various value principles to specific and existing pieces of legislation and regulation that may be applicable, as well as encourages innovation regarding the development of new standards, policies, and tools.

Australia

Currently Australia is closer to the US model than the EU model, although there is discussion that suggests a move towards the EU model of legislated regulation is possible.[16] Discussion papers and

[15] https://www.nist.gov/itl/ai-risk-management-framework

[16] See https://humanrights.gov.au/about/news/australia-needs-ai-regulation and https://www.corrs.com.au/insights/the-eu-ai-act-a-possible-direction-for-australian-ai-regulation

research papers have been produced and circulated in order to determine what approach will be best for Australia.[17] A cohesive Australia-wide response is not yet in place, although there is a set of voluntary ethical principles for AI technologies by the Australian Department of Industry, Science and Resources (DISR).[18]

New South Wales has a framework for AI technologies in place that includes a strategy, policy, set of ethical principles and associated guidance.[19] This includes an assurance framework designed to help organisations determine whether specific AI technologies should be implemented as part of a broader project.[20] The assurance framework is designed to flag potential areas of harm that may arise from using AI and provides high level guidance as to use of the AI technology. It should be noted that the set of ethical principles used by New South Wales differs from those used by the DISR (for example, there are five of them compared with the DISR's set of eight). The NSW set are also mandatory under Circular DCS-2020-04.

Various agencies across Victorian government (such as the Office of the Victorian Information Commissioner)[21] have positions and guidance on AI. The main areas of focus tend to be privacy, cyber security, and business innovation.[22] Currently there is no main centralised specific framework for addressing AI across Victorian government, although a new AI Hub / CoE is in development.

The below table provides a snapshot of AI approaches across Australasian jurisdictions regarding recordkeeping.

Jurisdiction	**AI Framework**	**Recordkeeping Guidance on AI**
Australia	https://www.industry.gov.au/publications/australias-artificial-intelligence-ethics-framework Voluntary compliance	https://www.naa.gov.au/information-management/managing-information-assets/types-information/information-management-current-emerging-and-critical-technologies
Australian Capital Territory	No	No
New South Wales	https://www.digital.nsw.gov.au/policy/artificial-intelligence Mandatory compliance with AI ethics; points to State Records Act in list of applicable legislation	No
New Zealand	https://www.digital.govt.nz/assets/Standards-guidance/Technology-and-	https://www.digital.govt.nz/standards-and-guidance/technology-and-

[17] See https://acola.org/rrir-generative-ai/ and https://consult.industry.gov.au/supporting-responsible-ai
[18] https://www.industry.gov.au/publications/australias-artificial-intelligence-ethics-framework
[19] https://www.digital.nsw.gov.au/policy/artificial-intelligence
[20] https://www.digital.nsw.gov.au/policy/artificial-intelligence/nsw-artificial-intelligence-assurance-framework
[21] https://ovic.vic.gov.au/privacy/resources-for-organisations/artificial-intelligence-understanding-privacy-obligations/
[22] See https://www.invest.vic.gov.au/opportunities/artificial-intelligence and https://www.vic.gov.au/a-future-ready-victoria/outcomes-victoria/digital-ready-public-sector and https://www.vic.gov.au/victorias-cyber-strategy-2021-introduction

	architecture/Generative-AI/Joint-System-Leads-tactical-guidance-on-public-service-use-of-GenAI-July-2023.pdf	architecture/interim-generative-ai-guidance-for-the-public-service/
Northern Territory	No	No
Queensland	No	No
South Australia	https://www.parliament.sa.gov.au/en/News/2023/07/11/03/37/Select-Committee-on-Artificial-Intelligence Research stage	No
Tasmania	No	No
Victoria	https://ovic.vic.gov.au/privacy/resources-for-organisations/artificial-intelligence-understanding-privacy-obligations/ In relation to Privacy specifically	https://prov.vic.gov.au/recordkeeping-government/document-library/approvalprocessespolicy-approval-processes-policy Mainly covers machine learning AI in relation to approval processes.
Western Australia	https://www.wa.gov.au/government/publications/artificial-intelligence-decision-making-initiative-2022	No

National Archives of Australia

National Archives of Australia (NAA) have added AI technologies to their *Current, Emerging and Critical Technologies* page with general advice on good recordkeeping that points out:

- Recordkeeping obligations.
- Information governance guidance.
- Managing information assets guidance.
- Guidance on keeping, disposing of and transferring records.
- Storage guidance.
- Additional resources pointing to external sites to assist with assessment and implementation guidance.

While the Australian Government has an *Artificial Intelligence Ethics Framework*,[23] currently NAA do not include it in their list of relevant resources.

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

Archives New Zealand

Archives New Zealand points to general advice for the public sector on Artificial Intelligence (with a focus on Generative AI) provided at Digital.Gov.NZ.[24] The advice was provided jointly by data, digital, procurement, privacy, and cyber security system leaders, including the Government Chief Digital Officer and Digital Government Leadership Group.

Archive New Zealand also points out the following:

- Information recording the use of AI pertaining to public and local authority records are subject to the *Public Records Act 2005* and the mandatory Information and records management standard.
- To meet the standard the following information must be retained and accessible until records are legally disposed (for example, transferred, destroyed or other disposal actions):
    - Evidence of what the AI application or system is.
    - What it was used for.
    - How it was used.
- Alongside the resources on Digital.govt.nz, Archives New Zealand are developing guidance specific to public and local authority records and use of AI to be published on our website in future.

State Records Authority New South Wales

New South Wales has a robust *Artificial Intelligence Framework*[25] which includes an assurance framework that mentions State Records legislation in the list of legislation applicable to AI.

State Records New South Wales do not appear to have any current specific guidance or information on public records and artificial intelligence.

Victoria

While Victoria does not yet have a specific Artificial Intelligence Framework and set of ethical principles, work is underway to establish an AI research and development hub (AI Hub) and centre of excellence (CoE) in Melbourne.

PROV has advice covering decision making and AI in the *Approval Processes Policy*.[26] This policy primarily addresses machine learning, including automated decision making and expert systems (AI technologies designed to analyse data and make decisions similar to a human). A challenge with these kinds of AI technologies is to ensure that there is no inherent bias in the data that has skewed the results to form a decision that will cause harm.

The Office of the Information Commissioner (OVIC)[27] has specific advice for AI on privacy requirements that covers multiple types of AI technologies and various points in the lifecycle of the

[24] https://www.digital.govt.nz/standards-and-guidance/technology-and-architecture/interim-generative-ai-guidance-for-the-public-service/
[25] https://www.digital.nsw.gov.au/policy/artificial-intelligence
[26] https://prov.vic.gov.au/recordkeeping-government/a-z-topics/policies
[27] https://ovic.vic.gov.au/privacy/resources-for-organisations/artificial-intelligence-understanding-privacy-obligations/

AI technology or system. This advice touches on recordkeeping in relation to the privacy requirements as part of the overall advice.

PROV does not yet have something that covers the challenges of generative AI, which are AI technologies that create content (for example, ChatGPT in the Microsoft 365 suite can take the minutes of a meeting for you, develop an agenda, provide a task list, write a draft report, gather and list relevant resources). A challenge with generative AI is that it can invent facts and write them up in a convincing fashion. This means that the minutes could contain incorrect information about what was actually said or omit actions, or that the draft report could pull information from sources that don't actually exist, and so on. Bias contained within the data sets used can also cause harm.[28]

While the *Approval Processes Policy* addresses questions of transparency in relation to approval processes, it does not cover the generation of accurate and trustworthy content, or any actions that may have taken place outside of an approval process. Additional guidance from PROV to cover these gaps is needed.

[28] The Algorithmic Justice League was created from direct experience of harm caused through AI and a desire to work towards the creation of ethical and transparent AI that work for everyone. Their website contains useful information about bias and harm in relation to AI technologies https://www.ajl.org/

An approach to AI and Recordkeeping in Victoria

As Victoria is establishing an AI Hub / CoE, it is likely that specific requirements for AI in the Victorian jurisdiction will be produced at some point. From what has already been put in place across the globe and within Australasia, it is likely that the approach will include a set of ethical principles and an assessment framework based around those principles.

PROV has in place advice on AI technologies in relation to approval processes, which would address AI technologies in relation to decision making as part of an approval process. In addition to this advice, the following is recommended for the Victorian jurisdiction in relation to recordkeeping:

- Advice on generative AI technologies and recordkeeping.
- Broader guidance on recordkeeping and AI technologies in order to address transparency and accountability regarding potential and actual bias and harm caused, and in relation to keeping full and accurate records.
- A list of existing external resources that Victorian government agencies may find useful.

Existing Useful Tools

There are a range of existing tools and associated guidance from across the globe that are designed to assess and document AI technologies. They can be used to assist with creating, capturing, and managing records created by or through the use of AI technologies in Victoria.

While the below tools were designed to address the situation within their specific countries and contexts, they are useful as stand-alone tools outside that context with a little tweaking (to ensure the tool aligns with the implementation context).

Algorithmic Transparency Report

A tool to document AI technologies based around the algorithms used.

The UK has developed a technical standard and guidance to enable public sector bodies to be transparent and pre-empt questions regarding what algorithms they may be using and how they may impact individuals. The associated guidance covers when to use the Standard, as well as how to complete it (including detailed examples for each section). Public sector bodies are to submit their completed reports to a central agency so that they are publicly available to view, accessible from a central hub.[29]

The standard is primarily a metadata template that captures specific information about algorithms used. It encourages the consideration of risks, impacts and accountabilities as part of the design and development phase of the tools. The report is to be completed during the design and development phase, and the report published during the pilot/deployment phase. The information required for the report from third party suppliers is considered to be at a general level which should not impact intellectual property. Legal advice and checking of contractual obligations may be needed.

It is not always a good idea to make the reports publicly available as sometimes that can be counter to the public good, so the guideline suggests to not make the information public if it would not be made available under Freedom of Information legislation. The guideline recommends an approach that is focused on releasing what can be released, with any sections not to be disclosed being redacted along with an explanation as to why.

Completion of the report can help to identify security risks with regard to the tool itself or with the information being recorded in the report. Where there is a security risk, the level of detail used in

[29] https://www.gov.uk/government/collections/algorithmic-transparency-recording-standard-hub

the report should be appropriate for public release (a secure and more detailed version could also be kept, should it be needed). Mitigation methods for the area of risk can also be determined and related to the report.

NIST AI RMF Playbook

A tool to aid analysis, identify information needed and document results across the AI lifecycle.

The US National Institute of Standards and Technology (NIST)has produced an *AI Risk Management Framework (AI RMF)*, and associated Playbook that is aligned with the Blueprint.[30] Released in January 2023, the framework centres the characteristics of trustworthiness and was intended for voluntary use, rather than being mandatory.

"The core of the AI RMF is composed of four functions: govern, map, measure, and manage. Each of the functions is then broken down into categories and subcategories. The govern function is intended to help cultivate a culture of risk management; the map function is intended to help recognize context and identify risks; the measure function is intended to help assess, analyze, or track risks; and the manage function is intended to help prioritize and act upon identified risks. The categories and subcategories break these functions down into numerous components, while the playbook provides actions, documentation guidance, and references for each subcategory. The AI RMF is designed to be applied in an iterative manner and used throughout the AI lifecycle."[31]

The Playbook[32] can be downloaded in multiple formats, including as a CSV file or MS Excel spreadsheet. This enables the tool to be aligned with organisation specific contexts, systems, and tools.

A Taxonomy of Trustworthiness

A tool to aid with identification and documentation of harms and biases.

The Taxonomy of Trustworthiness[33] is intended to compliment and support the NIST AI RMF. It analyses the landscape of trustworthy AI, is divided across the seven stages of the AI lifecycle, and provides 150 properties of trustworthiness.

The taxonomy includes questions to ask for each property of trustworthiness, connects each to the relevant NIST AI RMF subcategories and bolds those subcategories deemed a good place to start for easy reference.

capAI Conformity Assessment

A tool to assess for and document compliance with the EU AI Act.

The capAI tool[34] is a procedure for conducting conformity assessment of AI systems in line with the EU AI Act. It was produced by a group of University of Oxford researchers who describe it as follows:

"We have developed capAI, a conformity assessment procedure for AI systems, to provide an independent, comparable, quantifiable, and accountable assessment of AI systems that conforms with the proposed AIA regulation. By building on the AIA, capAI provides organisations with practical guidance on how high-level ethics principles can be translated into verifiable criteria that help shape

[30] https://www.nist.gov/itl/ai-risk-management-framework
[31] As described in Jessica Newman's *A Taxonomy of Trustworthiness for Artificial Intelligence* (https://cltc.berkeley.edu/publication/a-taxonomy-of-trustworthiness-for-artificial-intelligence/).
[32] https://airc.nist.gov/AI_RMF_Knowledge_Base/Playbook
[33] https://cltc.berkeley.edu/wp-content/uploads/2023/01/Taxonomy_of_AI_Trustworthiness.pdf
[34] https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4064091

the design, development, deployment, and use of ethical AI. The main purpose of capAI is to serve as a governance tool that ensures and demonstrates that the development and operation of an AI system are trustworthy – i.e., legally compliant, ethically sound, and technically robust – and thus conform to the AIA."

The tool consists of:

1. An internal review protocol for quality assurance and risk management
2. A summary datasheet
3. An external scorecard

The capAI procedure includes a discussion of the approach used by the EU AI Act and how it relates to an assessment framework, including a summary of the EU AI Act, determining level of risk, components of trustworthy AI systems, ethics-based auditing, and terminology used. The tool's review protocol follows the AI process flow. Each stage has a series of steps required, documentation in relation to that step, and the position of the person recommended to complete each step. The summary datasheet contains information about the AI technology or system being assessed. The prompts to develop an external scorecard containing summarised information for public consumption and to assist with transparency.

The tool is data focused, with a large section on evaluating the data to determine and mitigate bias, as well as managing data models, feedback mechanisms and so on.

Algorithmic Impact Assessment

A tool to quantify risk and determine the level of impact of AI technologies, primarily related to automated decision making.

The tool consists of an online questionnaire[35] and associated guidance including the *Directive on Automated Decision Making*[36].There are 51 risk and 34 mitigation questions, with assessment scores based on factors that include the system's design, algorithm, decision type, impact, and data. A score for each area is assigned, with the value of each question being weighted based on the level of risk. The raw impact score measures the risks of the automation, while the mitigation score measures how the risks of automation are managed.

The impacts of automating an administrative decision are classified into 4 levels, based on criteria of reversibility, and expected duration (automated decisions with little to no impact are reversible and brief, while those with a very high impact are irreversible and perpetual). Each impact level corresponds to a score percentage range. Impact levels determine the mitigations required under the *Directive on Automated Decision-Making*. Appendix C of the directive lists mitigation measures required for each of the 4 impact levels. The requirements are designed to be proportionate to the impact level. While the measures are intended to reduce the identified risks, their implementation does not alter the impact level assigned to a project.

[35] https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html
[36] https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai.html

Governing General Purpose AI

A tool to help understand unreliability, misuse and systemic risk re general purpose AI and document policy-based solutions.

Governing General Purpose AI[37] describes how general-purpose AI work as well as the kinds of risk they are prone to. Three risk categories of unreliability, misuse, and systemic risk are explored and broken down into subcategories. Each is described in relation to how it occurs, reference to case studies where harm eventuated as a result, and proposes potential harms that could result. The paper seeks to inform policy developers of what needs to be examined and addressed through policy and other means when using AI technologies to address and minimise harm.

NSW AI Assurance Framework

A tool to document AI technology implementation across the lifecycle of a project.

The *AI Assurance Framework*[38] is part of a suite of products from Digital NSW that includes a strategy and ethics policy. The Assurance Framework centres around a set of questions and associated response prompts that follow the lifecycle of a project. The questions are aligned with the five AI Ethical Principles that are described in the *Ethics Policy*[39] and includes elements for identifying risks and benefits as part of the assessment process. The response prompts include references to actions required or documentation needed, and point to other relevant NSW government bodies or products where relevant (such as privacy, security, and human rights impact assessments).

Explaining Decisions of AI

A tool to describe AI technology services, decisions, and processes so they can be explained to others.

The ICO / Alan Turing Institute guidance on explaining decisions of AI[40] aims to give organisations practical advice to help explain the processes, services and decisions delivered or assisted by AI technologies, to the individuals affected by them. It provides:

- detailed advice on how to develop and deliver different kinds of explanations depending on the outcome desired and the nature of the AI technologies used
- step by step advice through the explanation process including different scenarios and additional resources
- checklists for each step of what to consider
- how to build systems to enable explanations and the different challenges that may be faced depending on the type of AI technology concerned
- what kind of training implementers of the system will need to ensure that they can provide an explanation on the fly where needed.

The advice is data focused, and enmeshed within the technology used. It provides a good coverage of what would be needed for full and accurate records to be created and kept, when used to

[37] https://www.stiftung-nv.de/de/publikation/governing-general-purpose-ai-comprehensive-map-unreliability-misuse-and-systemic-risks
[38] https://www.digital.nsw.gov.au/policy/artificial-intelligence/nsw-artificial-intelligence-assurance-framework
[39] https://www.digital.nsw.gov.au/policy/artificial-intelligence/artificial-intelligence-ethics-policy
[40] https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/explaining-decisions-made-with-artificial-intelligence/

determine what the record would need to cover and what parts of the systems or processes at what points would need to be captured as a record.

Conclusion

The recommended approach for managing accurate and trustworthy records created by or through the use of AI technologies across Victorian government is as follows:

- That general advice be developed by PROV in the form of a topic page that covers:
    - what AI technologies are
    - what needs to be considered regarding creating, capturing, and managing records
    - how the general advice links with existing PROV advice, including the current *Approval Processes Policy*.
- That new policy advice regarding Generative AI technologies (including Large Language Models or LLMs) be produced to clarify what is required from a recordkeeping perspective with a focus on accuracy.
    - The existing *Approval Processes Policy* covers automotive and machine learning AI technologies that make and / or action decisions, and therefore focuses on ensuring transparency.
    - Generative AI technologies produce convincing and human-like content, such as minutes of meetings, reports, and artwork, based on what the user is looking for and have been known to invent content. The Generative AI Technologies guidance would therefore relate to the creation and capture of full and accurate records of business.
- That existing assessment frameworks and associated tools for documenting AI technologies across the globe be connected to the general AI guidance as useful external resources for Victorian government agencies.
- That the PROV advice be reviewed and updated once an approach from the AI Hub / CoE has become available.

Appendix One: Common Types of AI

A full list of AI types are included in the Glossary at Appendix Five.

Type	Description[41]	Records required to verify:
Automated Decision Making (ADM) Systems	use data to classify, analyse and make decisions that affect people with little or no human intervention	- that the decision was lawful - that the decision was made in line with appropriate procedures and lines of authorisation - that there was no bias involved or that bias was addressed appropriately - that the decision was based on accurate and current information
Expert Systems	use a knowledge base, inference engine and logic to mimic how humans make decisions	- that the decision was lawful - that the decision was made in line with appropriate procedures and lines of authorisation - that there was no bias involved or that bias was addressed appropriately - that the decision was based on accurate and current information
Generative AI	systems that produce code, text, music, or images based on text or other inputs	- that the content produced has been fact checked, and is accurate, current, and appropriate - that the content produced is in line with relevant legislation and regulation
Large Language Model (LLM)	a type of generative AI that specialises in the generation of human-like text	- that the content produced has been fact checked, and is accurate, current, and appropriate - that the content produced is in line with relevant legislation and regulation
Multimodal Foundation Model (MfM)	a type of generative AI that can process and output multiple data types (e.g. text, images, audio)	- that the content produced has been fact checked, and is accurate, current, and appropriate - that the content produced is in line with relevant legislation and regulation
Machine Learning Systems (MLS)	a broad set of models that have been trained on pre-existing data to produce useful outputs on new data	- that the content produced has been fact checked, and is accurate, current, and appropriate - that the content produced is in line with relevant legislation and regulation

[41] Definitions from: Solomon, L., & Davis, N., (2023); Australian Government, Department of Industry, Science and Resources (2023), *Safe and Responsible AI in Australia*, Department of Industry, Science and Resources

Natural Language Systems	models that can understand and use natural language and speech for tasks such as summarisation, translation, or content moderation	- that the content produced is authorised and valid - that the content produced has been fact checked, and is accurate, current, and appropriate - that the content produced is in line with relevant legislation and regulation
Robotic Process Automation	systems that imitate human actions to automate routine tasks through existing digital interfaces	- that the action was lawful - that the action was made in line with appropriate procedures and lines of authorisation - that there was no bias involved or that bias was addressed appropriately - that the action was based on accurate and current information
Virtual Agents and Chatbots	digital systems that engage with customers or employees via text or speech	- that the engagement content produced is authorised and valid - that the engagement content produced has been fact checked, and is accurate, current, and appropriate - that the engagement content produced is in line with relevant legislation and regulation

Appendix Two: Risks and Harms and their Mitigation

Risk / Harm	Description	Mitigation re Recordkeeping
Human Rights Risks / Harm due to Bias (Race, Gender, Economic, Age &c)	Occurs for various reasons, including the data set not being sufficiently diverse or having inherent bias that is magnified by the AI Examples include racism in predictive AI due to their being racism in the underlying data; FRS not recognising black faces as they have only been trained on white faces (see https://www.ajl.org/)	- approval process decisions captured and addressed in line with *Approval Processes Policy* - notification that AI is being used created/captured and managed along with content - create and capture records of monitoring and addressing risk/areas of bias - retain for duration of retention period and dispose of content lawfully.
Accuracy Risks / Harm due to Incorrect / inaccurate information – Fiction reported as Fact	Referred to as an 'hallucination'. Occurs for various reasons, including the language used by the human and the way the AI was taught to respond. Examples include references being invented by the AI, unverified content being referred to as fact by the AI (see https://www.usatoday.com/story/opinion/columnist/2023/04/03/chatgpt-misinformation-bias-flaws-ai-chatbot/11571830002/ and https://www.sify.com/ai-analytics/the-hilarious-and-horrifying-hallucinations-of-ai/)	- notification that AI is being used and what content is being generated by it created, captured and managed along with content - create and capture records of due diligence, e.g. fact checking and confirming whose intellectual property it is - retain for duration of retention period and dispose of content lawfully.
Transparency Risks / Harm due to Inability to explain AI Decisions and Actions	Occurs when it is unclear whether an AI was involved in making decisions or taking action, and/or how an AI was involved. (see https://www.uts.edu.au/sites/default/files/2023-05/HTI%20The%20State%20of%20AI%20Governance%20in%20Australia%20-%2031%20May%202023.pdf)	- notification that AI is being used and what content is being generated by it created, captured, and managed along with content - approval process decisions captured and addressed in line with *Approval Processes Policy*

Appendix Three: Good Recordkeeping and AI

Analysis of the various research, discussion, and other papers on AI technologies (see Appendix Six: References), their challenges, and applications, provides a set of common themes that are relevant for recordkeeping. They are as follows:

1. Acknowledge that AI technologies are being used, which ones, what they are being used for, and how they work throughout their lifecycles.
2. Provide a notification for individuals interacting with generative AI and other AI technology / applications along with information regarding opting out and request for review processes. Include information on any 'black box technologies' used.
3. Identify, describe, and document each AI system used across the organisation, including:
    a. the intended purpose and outcomes,
    b. desired benefits,
    c. type of AI used,
    d. context and scope of use,
    e. stage of implementation,
    f. sources of data,
    g. identified harms and risks,
    h. any controls or systems in place to mitigate risks, and
    i. other relevant information, such as what cannot be documented (for example, 'black box technologies' that do not disclose algorithms and/or other elements used to make decisions)
4. Document and report on how the system has worked over time (for example, using automated logging to record events)
5. Determine and document what content can be created and captured by AI technologies, what needs to be created and captured by a human being, and for AI generated content, at what points a human being is to be included.
6. Determine and document what actions can be undertaken by AI technologies and when a human being is to be included, confirm, or undertake the actions.
7. Be transparent and accountable regarding ownership and responsibility in relation to all types of AI used.
    a. Provide a clear attribution for (or transparent acknowledgement regarding issues attributing ownership of) text, images, sound, or video produced by generative AI.
    b. Establish appropriate oversight of and responsibility (including lines of delegation and accountability) for each AI system used that factor in AI system failures, overuse of AI, and malicious use.

c. Determine the legal requirements or obligations applicable to each system and carefully attribute legal liability for harms or errors to entities across the AI value chain, with effective safeguards placed at the most effective and appropriate points.

8. Establish an ongoing assessment and monitoring program that assesses AI technologies for trustworthiness.

    a. Use an assessment framework that is designed around a set of ethical principles to assess the AI technologies and systems from the idea/ design phase through to implementation and ongoing operations.

    b. Assess the AI technologies performance in relation to risk, bias, and harm and take steps to mitigate or otherwise address that risk, bias or harm.

    c. Document and report on the results as part of an ongoing monitoring and assessment program.

Appendix Four: Ethical Principles and Recordkeeping

The below table uses the set of mandatory ethical principles from the *NSW AI Ethics Policy*[42] that the Assurance Framework aligns with. The recordkeeping guidance column in the table below was drawn from the set of questions used in the *NSW AI Assurance Framework* and broadly tailored to the Victorian jurisdiction. It could potentially be developed for use as a general recordkeeping checklist. Its purpose is to illustrate at a high level how a set of ethical principles may help with determining what records of AI need to be kept.

Ethics principle	Recordkeeping considerations
Community Benefit AI should deliver the best outcome for the citizen, and key insights into decision making.	- Documentation to justify why the AI is being used and what it is used for - Evidence of alignment with relevant legislation, including the *Public Records Act 1973* - Connection with Privacy Impact Assessment – see OVIC - alignment with the Human Rights Charter – see Victorian Human Rights Commission (VHRC) - Documentation of how decisions are reached - Risk assessment documenting possibility of harms (eg inaccurate information / bias / IP breach) and their mitigation (if possible) or consequences (if not)
Fairness Use of AI will include safeguards to manage data bias or data quality risks, following best practice and Australian Standards	- Documentation showing results of assessment for accuracy, bias and associated mitigation - IP, copyright, and other data permissions / human rights charter alignment / Privacy Impact Assessment – see OVIC and VHRC - Evidence showing data integrity and associated analysis - Systems performance reporting
Privacy and Security AI will include the highest levels of assurance.	- Privacy Impact Assessment – see OVIC - Security Assessment - Human Rights Charter alignment – see VHRC - Data / Information Governance
Transparency Review mechanisms will ensure citizens can question	- Full and accurate records - evidence of how decisions were reached kept - what level of detail is possible - risk assessment and mitigation re harm / other impact - Evidence of consultation

[42] https://www.digital.nsw.gov.au/policy/artificial-intelligence/artificial-intelligence-ethics-policy

and challenge AI based outcomes.	- Open access information about AI - what is publicly available - Appeals process and associated documentation
Accountability Decision-making remains the responsibility of organisations and Responsible Officers	- Full and accurate records - evidence of how decisions were reached kept - what level of detail is possible - risk assessment and mitigation re harm / other impact - integrity of records - Chains of responsibility mapped - Appeals process and associated documentation / human in the system at point of authorisation

Appendix Five: Glossary

Types of artificial intelligence currently known are described below in alphabetical order and could be useful to include in PROV guidance. Please note that new types of AI will continue to be generated and used.

Term	Definition / Description
Algorithm	Automated instructions for a computer to perform a task or solve a problem.[43]
Application program interfaces (APIs):	A set of protocols to enable two software programs to communicate with one another, or for one program to run another program.[44]
Architecture	The design or structure of an AI model.[45]
Artificial intelligence (AI)	A collection of interrelated technologies used for problem solving and to complete tasks that would otherwise require human intelligence. [46] "…an engineered system that generates predictive outputs such as content, forecasts, recommendations or decisions for a given set of human-defined objectives or parameters without explicit programming. AI systems are designed to operate with varying levels of automation."[47]
AI model:	A program that has been trained on a dataset to recognise patterns (typically using artificial neural networks) or reason diagnostically or predictively (as seen in probabilistic graphical models).[48]
AI-Powered Robotics	Physical systems that use computer vision and machine learning models to move and execute tasks in dynamic environments. [49]
Artificial neural network	A type of machine learning consisting of a network of nodes that function analogously to the human brain.[50]
Automated Decision Making (ADM)	Systems that use data to classify, analyse and make decisions that affect people with little or no human intervention.[51] For example, to: • "make the final decision • make interim assessments or decisions leading up to the final decision • recommend a decision to a human decision-maker

[43] Bell, G., Burgess, J., Thomas, J., and Sadiq, S. (2023, March 24). *Rapid Response Information Report: Generative AI - language models (LLMs) and multimodal foundation models (MFMs).* Australian Council of Learned Academies.
[44] *Bell, G., Burgess, J., Thomas, J., and Sadiq, S. (2023).*
[45] *Bell, G., Burgess, J., Thomas, J., and Sadiq, S. (2023).*
[46] *Bell, G., Burgess, J., Thomas, J., and Sadiq, S. (2023).*
[47] Australian Government, Department of Industry, Science and Resources (2023)
[48] *Bell, G., Burgess, J., Thomas, J., and Sadiq, S. (2023).*
[49] Solomon, L., & Davis, N., (2023)
[50] *Bell, G., Burgess, J., Thomas, J., and Sadiq, S. (2023).*
[51] Solomon, L., & Davis, N., (2023)

	• guide a human decision-maker through relevant facts, legislation or policy • automate aspects of the fact-finding process which may influence an interim decision or the final decision. Automated systems range from traditional non-technological rules-based systems to specialised technological systems which use automated tools to predict and deliberate. "[52]
Expert Systems	Systems that use a knowledge base, inference engine and logic to mimic how humans make decisions. [53]
Facial Recognition Technologies	Systems that verify a person, identify someone, or analyse personal characteristics using face data drawn from photos or video. [54]
Foundation models	**"**Large AI-based models, trained on vast datasets, that can be applied to a variety of different tasks. Foundation models represent a paradigm shift in AI highlighting the phenomenal progression from algorithms (e.g. logistic regression), to architectures (e.g. transformers) to foundation models (e.g. GPT-3)."[55]
Generative AI	A type of AI model that can generate content such as text, images, audio and code, in response to user prompts. [56] Systems that produce code, text, music, or images based on text or other inputs. [57]
High-Risk AI Systems	"…in the light of their intended purpose, they pose a high risk of harm to the health and safety or the fundamental rights of persons, taking into account both the severity of the possible harm and its probability of occurrence..".[58] "The classification of an AI system as high-risk is based on the intended purpose of the AI system, in line with existing product safety legislation. Therefore, the classification as high-risk does not only depend on the function performed by the AI system, but also on the specific purpose and modalities for which that system is used."[59]
Large Language Model (LLM)	A type of generative AI that specialises in the generation of human-like text. [60]

[52] Australian Government, Department of Industry, Science and Resources (2023)
[53] Solomon, L., & Davis, N., (2023)
[54] Solomon, L., & Davis, N., (2023)
[55] *Bell, G., Burgess, J., Thomas, J., and Sadiq, S. (2023).*
[56] *Bell, G., Burgess, J., Thomas, J., and Sadiq, S. (2023).*
[57] Solomon, L., & Davis, N., (2023)
[58] EU AI Act s32
[59] EU AI Act 5.2.3. HIGH-RISK AI SYSTEMS (TITLE III)
[60] Australian Government, Department of Industry, Science and Resources (2023)

Machine Learning	The patterns derived from training data using machine learning algorithms, which can be applied to new data for prediction or decision-making purposes.[61] The development of models that can autonomously 'learn' from datasets and from inputs continuously [62]
Multimodal Foundation Model (MfM)	A type of generative AI that can process and output multiple data types (e.g. text, images, audio). [63] "…they use a wider range of information [than LLMs], including images, speech, numerical inputs and code, and they are trained on the relationship between the various inputs. Like LLMs, MFMs generate output based on learned patterns from the training input. Like LLMs, they also require tremendous computational power to train their models." [64]
Natural language processing	**"**An interdisciplinary branch of computer science, linguistics and artificial intelligence concerned with human–computer interaction and processing using human language."[65]
Natural Language Systems	Models that can understand and use natural language and speech for tasks such as summarisation, translation, or content moderation.[66]
Parallel processing	**"**Using multiple computing processors concurrently, enabling the processing of larger amounts of data in a shorter amount of time."[67]
Recommender Systems	Systems that suggest products, services or information to a user based on user preferences, characteristics, or behaviour.[68]
Robotic Process Automation	Systems that imitate human actions to automate routine tasks through existing digital interfaces.[69]
Technology stack	**"**The combination of technologies used to develop an application. May be used interchangeably with 'tech stack'."[70]
Transformer architecture	**"**A neural network that has the feature of learning context and parallel processing, enabling more powerful and faster models. This is due to a number of underlying advances, including an encoder-decoder structure."[71]

[61] Australian Government, Department of Industry, Science and Resources (2023)
[62] *Bell, G., Burgess, J., Thomas, J., and Sadiq, S. (2023).*
[63] *Bell, G., Burgess, J., Thomas, J., and Sadiq, S. (2023).*
[64] *Bell, G., Burgess, J., Thomas, J., and Sadiq, S. (2023).*
[65] *Bell, G., Burgess, J., Thomas, J., and Sadiq, S. (2023).*
[66] Solomon, L., & Davis, N., (2023)
[67] *Bell, G., Burgess, J., Thomas, J., and Sadiq, S. (2023).*
[68] Solomon, L., & Davis, N., (2023)
[69] Solomon, L., & Davis, N., (2023)
[70] *Bell, G., Burgess, J., Thomas, J., and Sadiq, S. (2023).*
[71] *Bell, G., Burgess, J., Thomas, J., and Sadiq, S. (2023).*

Trustworthiness	"The demonstrable likelihood that the system performs according to designed behavior under any set of conditions as evidenced by characteristics including, but not limited to, safety, security, privacy, reliability and resilience. In computer security, a chain of trust is established by validating each component of hardware and software from the bottom up.*"*[72]
UX	**"**User experience."[73]
Virtual Agents and Chatbots	Digital systems that engage with customers or employees via text or speech.[74]
Vision models	**"**A type of AI that can process visual information." [75]

[72] Newman, Jessica (2023), *A Taxonomy of Trustworthiness for Artificial Intelligence: connecting properties of trustworthiness with risk management and the AI lifecycle*, Center for Long-Term Cybersecurity UC Berkeley
[73] *Bell, G., Burgess, J., Thomas, J., and Sadiq, S. (2023).*
[74] Solomon, L., & Davis, N., (2023)
[75] *Bell, G., Burgess, J., Thomas, J., and Sadiq, S. (2023).G*

Appendix Six: References

The following references were used when drafting this paper:

Algorithmic Justice League (website dedicated to researching and raising awareness of the harms caused by AI and towards building a more equitable and accountable AI ecosystem), accessed August 2023 https://www.ajl.org/

Australian Government, Department of Industry, Science and Resources (2023), *Safe and Responsible AI in Australia*, Department of Industry, Science and Resources

Australian Government, Digital Transformation Agency (2023, July*). Interim guidance on generative AI for Government agencies*. DTA, Canberra, ACT. Accessed August 2023 https://architecture.digital.gov.au/guidance-generative-ai

Bell, G., Burgess, J., Thomas, J., and Sadiq, S. (2023, March 24). *Rapid Response Information Report: Generative AI - language models (LLMs) and multimodal foundation models (MFMs)*. Australian Council of Learned Academies.

Brennan, Jenny (2023), AI Assurance? Assessing and Mitigating Risks Across the AI Lifecycle. Ada Lovelace Institute, London, UK. Accessed September 2023: https://www.adalovelaceinstitute.org/report/risks-ai-systems/

Centre for Data Ethics and Innovation UK (2023, January). *Algorithmic Transparency Recording Standard - Guidance for Public Sector Bodies*. CDEI, Department of Science, Innovation and Technology, London, UK.

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.

Digital New South Wales, NSW Government (2022, March). *NSW Artificial Intelligence Assurance Framework*. Digital New South Wales, NSW Government. Accessed August 2023 https://www.digital.nsw.gov.au/policy/artificial-intelligence/nsw-artificial-intelligence-assurance-framework

Digital New South Wales, NSW Government (2022, March). *NSW Artificial Intelligence Ethics Policy*. Digital New South Wales, NSW Government. Accessed August 2023 https://www.digital.nsw.gov.au/policy/artificial-intelligence/artificial-intelligence-ethics-policy

Digital New South Wales, NSW Government (2022, March). *NSW Artificial Intelligence Strategy*. Digital New South Wales, NSW Government. Accessed August 2023 https://www.digital.nsw.gov.au/policy/artificial-intelligence/artificial-intelligence-strategy

European Commission (2020). *White Paper On Artificial Intelligence - A European approach to excellence and trust*. European Commission, Brussels, Belgium EU.

European Commission. (2021). *Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts*. COM/2021/206 final. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206

Floridi, Luciano and Holweg, Matthias and Taddeo, Mariarosaria and Amaya Silva, Javier and Mökander, Jakob and Wen, Yuni, *capAI - A Procedure for Conducting Conformity Assessment of AI*

Systems in Line with the EU Artificial Intelligence Act (March 23, 2022). Available at SSRN: https://ssrn.com/abstract=4064091 or http://dx.doi.org/10.2139/ssrn.4064091

Government of Canada (2023), *Directive on Automated Decision Making*, government of Canada, available August 2023 from https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai.html

Independent high-level expert group on artificial intelligence (AI HLEG) set up by the European Commission (2019). *Ethics Guidelines for Trustworthy AI*. AI HLEG European Commission, Brussels, Belgium EU.

Information Commissioner's Office and Alan Turing Institute (2023), *Explaining Decisions Made with AI*, ICO & Alan Turing Institute London UK. Accessed September 2023: https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/explaining-decisions-made-with-artificial-intelligence/

Maham, and Küspert, Sabrina (July 2023), *Governing General Purpose AI: A Comprehensive Map of Unreliability, Misuse and Systemic Risks*, Stiftung Neue Verantwortung, Berlin, Germany. Accessed August 2023: https://www.stiftung-nv.de/de/publikation/governing-general-purpose-ai-comprehensive-map-unreliability-misuse-and-systemic-risks

Newman, Jessica (2023), *A Taxonomy of Trustworthiness for Artificial Intelligence: connecting properties of trustworthiness with risk management and the AI lifecycle*, Center for Long-Term Cybersecurity UC Berkeley

Prud'homme, Benjamin and Régis, Catherine and Farnadi, Golnoosh (Editors) (2023), *Missing Links in AI Governance*, Published in 2023 by the United Nations Educational, Scientific and Cultural Organization (UNESCO). Accessed August 2023: https://www.unesco.org/en/articles/missing-links-ai-governance

Solomon, L., & Davis, N., (2023) The State of AI Governance in Australia, Human Technology Institute, The University of Technology Sydney

White House Office of Science and Technology Policy (2022, March*). Blueprint for an AI Bill of Rights: Making automated systems work for the American people*. White House Office of Science and Technology Policy, Washington, USA