Johniel Bocacao, Ali Knott, Andrew Lensen
Te Herenga Waka Victoria University of Wellington – 16 March 2026

Re-Engineering New Zealand Government Guidance for Trustworthy AI System Delivery: A Whitepaper

For government AI systems, the right fundamental laws that promote good government already enforce best practice in AI system design

Automated systems must be able to faithfully explain how and why they reached a decision

Staff must comprehensively keep records around how AI affected all aspects of an agency's affairs

Systems must only use people's data for the purpose it was collected and keep it secure

Automating statutory decisions and actions require explicit legislation

The main opportunity is streamlining all the different guidance that translates statutory obligations to operational instructions for AI deployers

GCDS guidance governs the use of personal data

- Algorithm Charter toolkit
- Five Safes
- DPUP
- Ngā Tikanga Paihere

GCDO guidance gives effect to mandatory technical standards

- Public Sector AI Framework
- Government Web Standards
- Responsible AI Guidelines
- Cloud risk discovery tool

The GCISO mandates cybersecurity practice

- PSR
- NZISM

A future New Zealand Artificial Intelligence Manual integrates technical guidance, coherently distinguishing between AI techniques and use cases

Common risks and considerations exist across all AI techniques

Common risks can manifest differently in different AI techniques

Different use cases have unique obligations, esp. in frontline and research

AI systems must be beholden to a strong normative standard promoting the trustworthy use of data agnostic of technology

One normative standard acts as the definitive lodestar for agencies developing a trustworthy data culture

DPUP already provides this: to focus on improving people's lives, and respecting, empowering and protecting them

"The use of data and Artificial Intelligence is the big opportunity of our time. We stand at the cusp of a digital revolution that has the power to transform the way our government serves New Zealanders. [...] I'd like to see the public service embrace the potential of AI."

– Minister for Digitising Government Hon Judith Collins to government chief executives in February 2025

The digital revolution sparked by new techniques in Artificial Intelligence (AI), integrated into user-friendly, general-purpose tools like ChatGPT, has swiftly transformed how people work, learn, and live. As businesses utilise AI to boost productivity and offer genuinely personalised customer experiences, government agencies in Aotearoa New Zealand ("the Government") are also adopting AI to improve how they serve the citizenry. However, just as businesses have employed traditional AI techniques for decades in narrow data-driven decision-making, the Government has been applying these methods since 2001 to inform decisions in offender management. Consequently, the Government has, over many years, contemplated what it means to develop effective, safe, fair, and accountable AI, especially given the mandatory, monopolistic nature of government actions and decisions, and the importance of maintaining citizen trust for effective governance.

Before ChatGPT and GenAI, the work culminated in the Algorithm Charter, voluntarily committing agencies to maintaining public trust in algorithms, including AI. While the Charter was principled, it lacked specifics and enforcement. The AIA toolkit added detail but, without enforcement, no agency has published a completed AIA.

In the age of GenAI, a parallel guidance framework has emerged for GenAI, while the Algorithm Charter work programme has not kept pace with the changes and the increase in the use of AI. Both now overlap significantly in scope and substantive guidance, while omitting useful substance from each other, resulting in duplicated interpretive effort when navigating them.

While experts[1] have rightfully called for further explicit regulation around the broader harms AI poses to New Zealand, this paper argues that – at least for government AI systems – the right fundamental legislation is already in place to enforce best practice in high-impact use of AI, while providing flexibility for lower-impact uses. Despite being older than modern computers, the Official Information Act 1982 sets an adequate floor for government transparency in the context of impactful automated decision-making. Instead, this paper calls for streamlining guidance across algorithms and all types of AI, given the technical universality of the major risks (and therefore mitigations) in any such system, and the legal universality of decisions and recommendations informed by technology. The legal principles are fit for purpose; unified guidance is all you need.

In addition to universal challenges, such unified guidance must consider both the unique challenges within certain types of algorithms and AI and the unique challenges within the governmental context in which these systems are deployed. I propose a taxonomy that provides sufficient depth to be useful in governance.

---

[1] *Andrew Lensen, Christopher McGavin, Cassandra Mudgway et al., Sep 2025. A call to the NZ Parliament to regulate AI. https://regulateai.nz*

About the Author

Johniel Bocacao has held multiple roles in the data and policy space across government over the last five years. He currently holds a *Postgraduate Certificate in Science* in Artificial Intelligence, with academic experience spanning the design of databases, data products, computer games and algorithms; software engineering practice; and offensive, defensive and organisational cybersecurity. He is working towards a *Master of Science* in Artificial Intelligence. This whitepaper provides an overview of the actionable conclusions from the thesis.

His current role at ACC has informed his perspective on the use of AI in an operational setting. With modern technology, security, privacy and data systems and data science capability, ACC leads in the careful yet innovative design and adoption of AI in the public service. His specific role in evidence and analytics across multiple organisations has informed how he sees AI use in a policy setting: delivering products in, and promoting the use of, predictive analytics to unlock a powerful evidence base for both public and private firms to make better decisions.

AI Use Statement

The use of AI in developing this whitepaper and the underlying thesis was limited to literature discovery and sentence-level edits for semantic accuracy, particularly for legal wording. Generative AI tools were used for creating iconography due to the bespoke requirements of the concepts being illustrated. The protracted process of realising my exact requirements – hitting the increased limit for a Gemini Pro subscriber – gives me hope that designers will not be losing their jobs any time soon.

Table of Contents

1. Rules and best practice in NZ for algorithm and AI system delivery are currently fragmented across different laws and guidance

Deployers of AI systems in government face a complex maze of laws, principles, standards, and guidance to deploy AI safely and effectively. This approach can be advantageous, as broad rules address long-lasting risks or support enduring objectives – rules relevant regardless of context or technology. Specific rules are created by specialised entities with the expertise to manage

certain risks or achieve certain objectives. Figure 1 depicts this guidance ecosystem as a layered pyramid, with higher-level instruments becoming more technically specific, while guidance in instruments below it remains applicable. Following guidance at all levels is essential to maximising the benefits.

This fragmentation becomes problematic when each specialist entity develops its own framework, toolkit, or checklist for AI, often with overlapping elements. This issue worsens for implementing agencies, which must interpret the various rules and tools and incorporate these frameworks into their policies. While this integration can help tailor governance to each agency's operational environment, a lack of a common scope and framework leads to inconsistent recordkeeping across agencies. This inconsistency makes it harder for the public – including academia and the media – or even a government system leader, to understand AI use across the government.

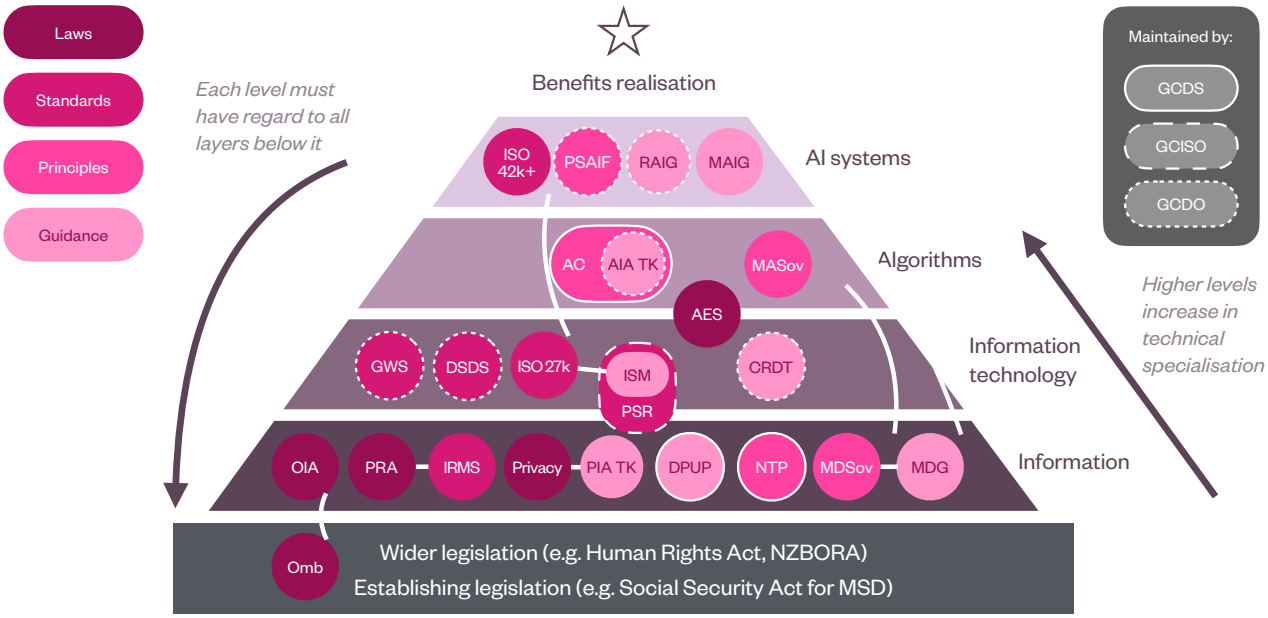


Figure 1: Current hierarchy of laws, standards, principles, and guidance that inform the use of information, information technology, algorithms, and AI systems, coloured by the type of instrument, and outlined if the instrument is maintained by a system leader. Following guidance in all layers is essential to fully realise its benefits.

1.1. Generic legislation mandates broad obligations on the use of AI, like transparency, privacy and information provenance

As previously discussed, generic legislation establishes lasting rules that persist regardless of context or technology. Three such acts serve as technology-neutral mechanisms that address specific risks through principles or rights-based regulation. The Official Information Act 1982 mitigates government opacity and promotes good decision-making and behaviour through freedom of information mechanisms (Box 1.1.1). The Privacy Act 2020 mitigates unauthorised and harmful collection, use, and disclosure of personal information, which infringes on individuals' rights to privacy (Box 1.1.2). The Public Records Act 2005 mitigates the erosion of accountability caused by poor and incomplete record-keeping (Box 1.1.3). Together, this ecosystem provides a stable regulatory baseline for democratic accountability. This baseline endures regardless of the technology in use and therefore applies to AI systems.

Box 1.1.1. Official Information Act 1982 (OIA)

Promotes government transparency and accountability, requiring all public service agencies to provide anyone in New Zealand with information the government holds. Sections 22 and 23 promote good government decision-making by providing anyone in New Zealand (including businesses) the right to know, respectively:

a) **How** a decision (including recommendations) was made: the **right** to access rules affecting decisions that can affect **anyone's** rights or interests. Any document (including electronic information) that contains such rules can be requested. [a]
b) **Why** a decision (including recommendations) was made: the **right** to access the reasons for decisions that affect the **requestor's** rights or interests. [b]

The Act specifies valid reasons for withholding this information. s6 outlines conclusive withholding grounds, while s9 outlines other withholding grounds that must outweigh the public interest in making information available. As these are **rights**, fewer s9 withholding grounds apply to rules/reasons requests. s18, which outlines grounds for refusing requests, do not apply to rules/reasons requests.

Requestors affected by general decisions, such as policy decisions, cannot use s23 to determine the reasons for such a general decision where it does not uniquely affect their own rights or interests. This information may be requested under a normal section 12 request, subject to other potential withholding grounds, such as free and frank opinions. [b]

A response to a s23 request must include:

1. the findings on material issues of fact (the data, evidence and inferences used to support a decision)
2. where those findings originated (not necessarily the documents themselves, but it is more administratively efficient to include them if they are requestable)
3. a logical explanation of the reasoning that links such facts to the conclusion, not just the conclusion itself – an approximate or post-hoc explanation is insufficient

These provisions are closely coupled with that of the **Ombudsman Act 1975** enabling the Ombudsman to investigate most government actions or decisions as contrary to law; unreasonable, unjust, oppressive, or improperly discriminatory; a mistake; or wrong.

For AI use in government, OIA ss 22 and 23 serve as the most significant constraint. Agencies that deploy AI that cannot faithfully explain personally impactful decisions deprive people of their right to information contravening the OIA. Unexplainable AI may still be used outside of making conclusive personally impactful decisions. Section 5 elaborates on the different degrees of compliance required for different uses of AI systems.

Further reading: [a] Requests for internal decision making rules: A guide to section 22 of the OIA and section 21 of the LGOIMA - Ombudsman New Zealand - May 2019 and [b] Requests for reasons for a decision or recommendation: A guide to section 23 of the OIA and section 22 of the LGOIMA - Ombudsman New Zealand - May 2019 and [c] AI systems and OIA section 22 and 23 - Ombudsman query - Johniel Bocacao - February 2026

Box 1.1.2. Privacy Act 2020

Promotes the human right to privacy by requiring any holder of a person's information (not just a government agency) to follow these information privacy principles (IPPs), notwithstanding exceptions elsewhere in the Act.

- Collection of personal information has a lawful and necessary purpose (1), is collected from the individual where possible (2) who understands why it is collected (3), and is done in a lawful and reasonably unintrusive manner (4)
- Personal information is stored and secured against loss, and unauthorised access (5)
- Individuals have the **right** to access (6) and correct (7) information about them.
- Information is checked as accurate before use or disclosure (8)
- Information is not kept longer than necessary (9) nor used (10) or disclosed (11) for anything beyond the original purpose of collection.
- Information can only be disclosed to foreign entities who comply the Act or provide similar safeguards (12)
- Unique identifiers are only generated when necessary for an entity's functions, and cannot be reused by another entity (13)

This Act establishes the Privacy Commissioner, who may issue guidance (e.g. [a]), tools like the Privacy Impact Assessment, and legally binding codes of practice, outlining different standards (may be more or less stringent) for specific types of information, such as the Health Information Privacy Code applicable to all health agencies.

Wherever AI systems in government use personal information, the Privacy Act's principles prompt deployers to consider several questions:

- Does the information used come directly from the person it is about (2)? Is AI being used to generate predictions or assumptions, or draw on public sources, rather than collected from the person directly?
- Does the individual know their information may be used by (or used to train) AI tools (3)?
- Is personal information protected against unauthorised access, particularly where publicly available AI tools risk data leakage and other unintentional data memorisation (5)?
- Can individuals access and correct data being used to train an agency's predictive models (6-7)?
- Can individuals correct (7) or challenge the accuracy of a prediction or other AI-derived output (8)? Such AI-derived outputs itself are still considered personal information under the Privacy Act.
- Is the use of personal information in AI directly related to the original purpose of its collection (10)? Could it be reused by the AI provider, particularly publicly available AI tools?

Further reading: [a] Artificial intelligence and the Information Privacy Principles - Privacy Commissioner - Sept 2023

Box 1.1.3. **Public Records Act 2005 (PRA)**

Promotes comprehensive recordkeeping within government by regulating how all information about an agency's "affairs" is managed within them.

All information generated by a "government office", including AI-generated information, is considered a public record. Section 17 mandates that agencies "create and maintain full and accurate records of its affairs". Section 27 enables Archives New Zealand to set mandatory recordkeeping standards. The current standard has three principles, the third of which ("Information and records are well managed") elaborates on the obligations on agencies:

- Recordkeeping occurs as part of normal business practice
- Records must be reliable and trustworthy
- Records must be protected from unauthorised access, alteration or loss

Wherever AI is used in government, the use and impact of AI on an agency's affairs and information needs to be documented as part of routine recordkeeping. Archives New Zealand provides a checklist tool to promote compliance by recording what the AI system is and will be used for.

Further reading: Artificial intelligence and public and local authority records - Archives New Zealand

1.2. Automating statutory decisions and actions must be legislated

Establishing legislation, the laws that set up an entity and determine its mandate, may include provisions that enable Automated Electronic Systems (AES) to perform actions specified in legislation that only specified authorised people (typically the chief executive) can do, as if such an authorised person performed the action themselves. While the term is not defined in law, it can reasonably be understood to include any algorithm or AI (as defined in Section 4) that makes decisions without human intervention. AES provisions typically follow this pattern:

1. Arranging the use of an AES
    - Outlining which statutory roles and actions the system can substitute
    - System must be able to perform the actions "with reasonable reliability"
    - A human alternative must always be available to perform the action instead
    - Allowing systems with components outside New Zealand to be used
    - Mandating consultation with the Privacy Commissioner
2. Describing the effect of the use of the system
    - Treating its action as done properly, as if it were done by the specified person
    - If its actions are clearly wrong, they may be done by people
3. Describing offences against an AES if applicable
    - Failure to comply with requirements or directions from an AES
    - Obstructing, hindering, damaging or impairing an AES
    - Deceiving or withholding information from an AES
    - Covering legislation-specific offences, e.g. manipulating offending goods seized by an AES without its permission.

This pattern is not in legislation where powers are conferred on entities rather than specified individuals. For example, the Accident Compensation Corporation (ACC) – a Crown agent with well-publicised automated decision-making – is not empowered to use AES. Most of its administrative decisions are conferred on "the Corporation", not on a chief executive or a qualified person. ACC may need specific AES provisions if it wishes to automate complicated decisions that must be made by qualified assessors, such as rehabilitation support needs.

Explicit legislation is required to say the system's action counts as a specified person's action, if the legislation requires a specified person to perform an action to be automated.

1.3. Agencies may be given the power to enact secondary legislation on their own

Two examples of secondary legislation have already been mentioned:

- Privacy Act codes of practice, issued by the Privacy Commissioner
- Information and records management standard, issued by Archives New Zealand

Secondary legislation holds the same authority as any other Act of Parliament. However, it is usually drafted by the agency responsible for administering the Act. The Cabinet Manual describes secondary legislation as addressing matters of detail, technical issues, or those likely to require frequent changes. Therefore, this regulatory route is preferred for giving effect to generic principles (like those in the Acts mentioned) by clarifying obligations at a technical level. For the Privacy Act 2020, secondary legislation can also modify the Act's obligations in specific situations. For instance, the Health Information Privacy Code clarifies when health privacy protections are stronger (e.g. parents have less access once a child is 16 or older) or weaker (e.g. legal representatives of deceased or incapacitated individuals may access information).

AI systems must also follow any secondary legislation that applies to the context it is being used, or data it is using.

1.4. Government-wide **standards and guidance** provide all-of-Government best practice in particular areas

The Public Service Act 2020 enables the designation of a specific agency as a system leader to "lead and co-ordinate best practice in a particular subject matter area" across all government sectors. This leadership is realised through setting standards or guidance in such areas. Standards are applicable to public service agencies, while guidance applies to any Crown organisation. These standards and guidance were issued before the 2020 act, with varying levels of enforceability. Some are mandated by Cabinet directives or ministerial instructions; others are voluntary best practice. The following section describes relevant system leaders whose guidance influences the use of AI systems as of December 2025.

Box 1.4.1. Government Chief Digital Officer (GCDO)

Coordinates the development, procurement, assurance and use of digital systems across government. Operated out of the Department of Internal Affairs until April 2026.

- Maintains guidance and tools that facilitate best practice design and implementation of digital systems, including AI tools which are typically cloud-hosted Web interfaces.
  - Information sharing standard: facilitating protection of personal information held by government when systematically sharing or collecting information from a third party outside government. Mandated under s57 Public Service Act 2020.
  - Digital Service Design Standard: supporting the government to design and provide public services that are easily accessible, integrated, inclusive and trusted by all New Zealanders. Applies to public-facing or inter-agency digital services undertaken by external third parties.
  - Government Web Standards: usability and accessibility standards for both internal and public facing web interfaces. Mandated by Cabinet direction.
  - Risks assessment for public cloud services: helps identify risks and security controls to consider when using a public cloud service. Optional tool for assisting in mandatory risk assessments.
- Maintains guidance and frameworks for the safe and responsible use of AI
  - Public Service AI Framework (PSAIF) provides a strategic vision and principles for adopting AI responsibly across the public service.
  - Responsible AI Guidance (RAIG) for the Public Service: GenAI (RAIG-PSG) guides agencies in the development and use of GenAI systems, linked to the OECD AI Principles. RAIG is intended to be a series, for example the Ministry of Business, Innovation and Employment (MBIE) has developed RAIG for businesses.
- Mandates coordinated procurement of common digital products and services
  - ICT Common Capability contracts enable procurement of a commonly used technology once on behalf of multiple or all government agencies.
  - Assurance Services Panel convenes trusted providers of quality assurance services that agencies are required to use for any digital investment they make.
- Coordinates digital transformation across the public service through the Service Modernisation Roadmap. AI-related initiatives include:
  - All-of-Government AI reference architecture to guide the design and implementation of government's AI solutions, promote best practices and accelerate development across the public service.
  - Two-year roadmap to FY27/28 for accelerating use of government AI solutions

The GCDO is the key leader responsible for standards and guidance in AI systems and tools whose development is contracted out, as with most digital systems in government.

Box 1.4.2. Government Chief Information Security Officer (GCISO)

Issues mandatory security requirements to government agencies. These functions are hosted by the National Cyber Security Centre (NCSC) which provides guidance and intelligence for cybersecurity threats to public and private organisations. Both are operated out of the Government Communications Security Bureau.

- The GCISO maintains the Protective Security Requirements (PSR), expectations for managing all aspects of an agency's security: physical, personnel and information security, and security governance. Mandated by Cabinet Directive CAB MIN (14) 39/38.
  - New Zealand Information Security Manual catalogues essential controls and additional recommended controls.
- The NCSC works with its allied counterparts to issue non-binding joint guidance on secure AI use for any NZ organisation to consider:
  - Guidelines for secure AI system development, December 2023. Provides four high level guiding principles to frame secure AI system development: secure design, secure development, secure deployment, and secure operation & maintenance.
  - Engaging with artificial intelligence, January 2024. Outlines common challenges around AI and potential mitigations for model data poisoning, input manipulation, hallucination, privacy concerns, model stealing.
  - Deploying artificial intelligence (AI) systems securely, April 2024. Expands on the secure deployment and secure operation & maintenance of the December 2023 guidance, and builds on mitigations identified in the January 2024 guidance.
  - Artificial intelligence (AI) data security, May 2025. Provides guidance for securing data used during AI development and deployment.
  - Principles for the Secure Integration of Artificial Intelligence in Operational Technology, December 2025. Prescribes technical considerations around the use of AI in safety-critical equipment and facilities, such as electricity infrastructure management, hospital operation, traffic management, police and correctional facility security.

The GCISO's security mandates necessarily applies to any government AI model or system, including those developed in-house. The wider NCSC provides internationally endorsed best practice around securing AI systems.

Box 1.4.3. Government Chief Data Steward (GCDS)
Promotes accessibility, reliability and ethical use of data in government. Operated out of Statistics New Zealand.

- Maintains policies that promote trusted use of data and algorithms (including AI) in government, all of which are voluntary.
  - Data Protection Use and Policy (DPUP) outlines values, behaviours and practice beyond legal obligations to promote respectful, culturally considerate and transparent collection and use of personal data.
  - Algorithm Charter ("the Charter") is signed by individual agencies to signal how they design and implement algorithms in a transparent, people-focused, legally compliant, and human-overseen manner.
  - Algorithm impact assessment (AIA) toolkit helps agencies operationalise their Charter commitments, providing a specific threshold for systems in scope of the Charter, and specific guidance at each stage of development.
- Outside of the GCDS, StatsNZ is responsible for the maintenance of systems that securely integrate and match data from different agencies. Given the sensitivity around massively integrated data, policies surrounding their use are strictly enforced.
  - Five Safes framework conditions access to integrated data given: safe people, safe projects, safe settings, safe data, and safe output.
  - Ngā Tikanga Paihere outlining ten tikanga Māori concepts that promote culturally appropriate, good faith engagement with Māori and use of Māori data.

The GCDS provides voluntary guidance for the trustworthy, ethical and culturally appropriate use of government data inputted in AI systems and tools, and in the development of their own AI models.

1.5. Government agencies use international standard risk methodologies, and new standards are emerging to facilitate AI system risk management

The PSR requires agencies to follow ISO 31000 to manage risk, a widely practiced international standard. ISO 31000 does not impose a checklist exercise but actively involves organisations in identifying specific risks, analysing their impact, assessing risk acceptability, and deciding on appropriate actions (accept, transfer, mitigate, avoid).

The PSR also references the ISO 27000 family of standards, which provides specific guidance on information security risk management and is already widely used by government agencies. A new family of ISO standards is emerging for the governance and risk management of AI systems, mirroring the structure of the ISO 27000 family. There are gaps, as shown in Table 1, but all listed AI system standards are officially published with full ISO consensus.

Notably, early-adopter government agencies like ACC have opted to adopt the American federal government's standard for AI risk management, given their reliance on the same American standards for cybersecurity and privacy.

Function of standard	ISO 2700x Information system standards	ISO AI system standards
Definitions for shared vocabulary	ISO 27000	ISO 22989
Management system requirements (risk governance spine)	ISO 27001	ISO 42001
Catalogue of risk control guidance	ISO 27002	*No definitive equivalent, some in ISO 24027-29*
Measurement and metrics	ISO 27004	*No equivalent*
Tailoring ISO 31000 (risk management methodology) to a specific domain	ISO 27005	ISO 23894
Impact and control assessment	ISO 27008	ISO 42005

Table 1: Functional alignment between ISO information system and AI system standards

Mandatory security requirements are largely based on international standards. Agencies can confidently adopt emerging recognised international standards for AI risk management as future requirements are likely to closely align with them.

1.6. Frameworks help affirm Māori **sovereignty over their data, and algorithms** developed from their data

Iwi Māori retain tino rangatiratanga over all their taonga, as guaranteed in Article 2 of Te Tiriti o Waitangi. This rangatiratanga extends to any data about or from Māori people and culture, as outlined in Wai 2522[2]. Te Mana Raraunga, a network of Māori experts in data, has devised principles (which are currently not mandated across government) to promote practices that affirm Māori data sovereignty, as outlined in Table 2.

Rangatiratanga: Right to control and access Māori data and data ecosystems, storing data here enhances control	**Whakapapa**: Metadata should be accessible to provide context, origin, provenance. Methodologies should prioritise Māori aspirations	**Whanaungatanga:** Individual rights to data and privacy are balanced with those of the collective
Kotahitanga: Empower Māori to derive individual and collective benefit	**Manaakitanga:** Respect the dignity and consent of Māori, avoid deficit-based data practices	**Kaitiakitanga**: Enable Māori to protect their data at all parts of the lifecycle (e.g. storage, transfer, disclosure) according to tikanga

Table 2: Principles of Māori Data Sovereignty. From Te Mana Raraunga, October 2018. https://www.temanararaunga.maori.nz/principles-of-maori-data-sovereignty

[2] *Waitangi Tribunal, March 2023. The Report on the Comprehensive and Progressive Agreement for Trans-Pacific Partnership. https://www.waitangitribunal.govt.nz/en/news/tribunal-releases-report-on-cptpp*

This framework was extended by Brown et al. (2024) to promote the sovereignty of algorithmic systems (not just the algorithm itself, but the inputs, output interpretation, design, and ongoing management), informed by data about or from Māori. In addition to the obligations above regarding data used by algorithms, the framework also offers further algorithm-specific guidance for each principle, as outlined in Table 3.

Rangatiratanga: Right to control development and use of algorithmic systems	**Whakapapa**: Māori understand of all aspects of the algorithm system	**Whanaungatanga:** Right to challenge its output or outcome, system owners are accountable to Māori
Kotahitanga: As above, for algorithmic systems	**Manaakitanga:** Consider individual and collective privacy throughout the entire operation of the system	**Kaitiakitanga**: As above, for algorithmic systems

Table 3: Principles of Māori Algorithmic Sovereignty. From Paul T. Brown, Daniel Wilson, Kiri West, Kirita-Rose Escott, Kiya Basabas, Ben Ritchie, Danielle Lucas, Ivy Taia, Natalie Kusabs, Te Taka Keegan, April 2024. Māori Algorithmic Sovereignty: Idea, Principles, and Use. https://datascience.codata.org/articles/10.5334/dsj-2024-015

Te Kāhui Raraunga, under the direction of the Data Iwi Leaders Group, have devised Māori Data and AI Governance frameworks for use within the public service. The Māori Data Governance model implements the six principles from Te Mana Raraunga. Five values frame the directives under the eight pou, which categorise actions that give effect to the values, as shown in Table 4.

Vision: Tuia te korowai o Hine-Raraunga / Data for self-determination				
Values				
Nurture data as taonga	Use data for good	Put iwi Māori data in iwi Māori hands	Be accountable	Decolonise data systems
Pou (Pillars)				
1. Capacities and workforce development	2. Data / **IT** infrastructure	3. Data collection / **AI data generation**	4. Data protection	
5. Data access, sharing and repatriation	6. Data use and reuse / **for AI implementation**	7. Data / **AI** quality and system integrity		
8. Data classification				

Table 4: Māori Data Governance Model (with Māori AI Governance Model overlaid in bold). From Tahu Kukutai, Kyla Campbell-Kamariera, Aroha Mead, Kirikowhai Mikaere, Caleb Moses, Jesse Whitehead and Donna Cormack, 2023. Māori data governance model, Te Kāhui Raraunga, https://www.kahuiraraunga.io/maoridatagovernance and Te Kāhui Raraunga, 2025. Māori Artificial Intelligence Governance Framework. Contextualised advice for AI use, extending the Māori data governance model, https://www.kahuiraraunga.io/maoriaigovernance

Much like the NZISM outlines how to comply with the PSR's principles, the Māori Data and AI Governance Models outline requirements that affirm the principles of Māori data sovereignty that arise from Te Tiriti o Waitangi.

1.7. Agencies set their own **departmental policies and procedures to clearly establish accountability for following laws, standards and guidance**

These laws, standards, and guidance ultimately have operational effect when embedded in an agency's policies and enforced by the chief executive through managerial consequences. Departmental policies define legal obligations, set expectations, and tailor standards to their operating environments, serving as the most effective means of implementing and holding agencies accountable for complying with laws, standards, and guidance.

Most agencies with AI policies focus mainly on GenAI tools: authorising only approved tools, forbidding the use of sensitive information, raising awareness of risks, mandating training, and defining roles and responsibilities. Agencies that have published departmental policies for AI use include:

- Ministry of Social Development: predates generative AI but exemplifies mature policies and – more importantly – procedures for algorithm and AI development. Their ADM Standard[3] outlines baseline requirements for any algorithm that makes decisions, supplemented with technical guidance for more complex systems within their Model Development Lifecycle[4].
- Courts of New Zealand[5]: provides specific guidance for specific roles, with the same principles throughout, and plain English explanations of the limitations of GenAI and why the risks exist. Notably, the use of GenAI need not be disclosed unless requested by the courts.
- Land Information New Zealand[6]: anonymise names when discussing personal scenarios with AI, disclosing AI use for externally released work, work with Treaty partners if Māori data is involved, or Māori interests may be affected in the use of AI.
- Accident Compensation Corporation[7]: mandating human oversight throughout use, actively protecting taonga Māori in accordance with Treaty commitments, prescribing unacceptable use of GenAI (e.g. for client care, for resourcing decisions, with personal, health or confidential information).
- New Zealand Police[8]: mandating labelling to disclose where GenAI is being used, external vendor disclosure of GenAI use, requiring Tier 2 approval for urgent unapproved uses,

---

[3] *Ministry of Social Development, 2022. About MSD's Automated Decision-Making Standard. https://www.msd.govt.nz/about-msd-and-our-work/work-programmes/initiatives/phrae/adm-standard.html*

[4] *Ministry of Social Development and Nicholson Consulting, October 2021. Model Development Lifecycle. https://www.msd.govt.nz/about-msd-and-our-work/work-programmes/initiatives/phrae/model-development-lifecycle.html*

[5] *Courts of New Zealand, December 2023. Guidelines for use of generative artificial intelligence in Courts and Tribunals. https://www.courtsofnz.govt.nz/going-to-court/practice-directions/practice-guidelines/all-benches/guidelines-for-use-of-generative-artificial-intelligence-in-courts-and-tribunals*

[6] *Land Information New Zealand, December 2023. Artificial Intelligence (AI) Use Policy. https://www.linz.govt.nz/sites/default/files/2024-05/DOIA%2024-261%20Artificial%20Intelligence%20Use%20Policy%20signed.pdf*

[7] *ACC, August 2023. Generative AI Models and Services Policy. https://www.acc.co.nz/assets/oia-responses/Documents-relating-to-the-algorithm-charter-of-aotearoa-oia-response-gov-031495.pdf*

[8] *New Zealand Police, September 2025. Acceptable use of Generative AI. https://www.police.govt.nz/about-us/programmes-and-initiatives/police-use-emergent-technologies/generative-ai*

prohibiting fully automated (no human in the loop) GenAI, prohibiting use of AI in court, requiring audit logs for complete recordkeeping.

Departmental policies are ultimately how standards and guidance are given effect: as expectations on staff relying on internal accountability. They are also an opportunity to tailor policy around what agencies do and how they operate.

1.8. Action: Establish a New Zealand Artificial Intelligence Manual (NZAIM) to unify technical guidance in implementing trustworthy government AI systems

The right laws and principles for promoting trustworthy AI system delivery are already in place. The OIA's transparency requirements effectively bind agencies to thoroughly consider best practice – offered but not mandated by other guidance instruments – when designing transparent AI systems trusted to make decisions independently. Other best practice around fairness and non-discrimination are promoted through recognising the ex-post risk of an investigation finding AI decisions to be unlawful, mistaken or discriminatory.

However, the current guidance ecosystem is missing a keystone instrument that provides a single, authoritative source of technical guidance on controlling risks and promoting the norms identified by other relevant instruments.[9] To reduce existing fragmentation, this unified guidance should organise all the **objectives** of each guidance instrument and mandate or recommend technical **controls,** along with their **rationale**. This approach is already employed by the NZISM, which provides agencies with flexibility around specific controls it recommends but does not mandate, requiring agencies to document and accept the residual risk of non-use. These controls are supported by a structured certification and accreditation process standardising system assurance. This approach has streamlined cybersecurity practice in government and transparently models best practice for non-government entities.

An NZAIM can not only promote these outcomes for government AI but also for private AI by providing a consumer signal that accredits, rather than outright regulation that compels, best practice. The GCDO has committed to an AI assurance model and toolkit by 2027, as well as a separate deliverable on safety certifications. Like the NZISM, an NZAIM would encompass both.

The NZAIM would catalogue gold standard technical controls for AI risk management in government, helping technical practitioners give effect to all the different guidance instruments, and model best practice for the private sector.

[9] *It may be argued that the AIA toolkit fulfils such a role. It is primarily a risk assessment framework and only prescribes high-level controls, not to the same technical depth as the NZISM.*

2. The Algorithm Charter has yet to develop into the world-class framework that it was set to become

I argue that, at the time, Minister James Shaw rightfully described the Algorithm Charter as a world first. Back then, only the Government of Canada (GC) had comparable provisions in this space. The GC's Directive on Automated Decision Making (DADM) offers less detailed but more practical guidance for managing automated algorithms. It uses the same concrete language as the enterprise processes that operationalise it, giving agencies actionable advice:

- Like the Charter, the DADM requires an explanation of decisions. However, the DADM goes further by outlining what a sufficient explanation looks like
- Like the Charter, the DADM requires human oversight of decisions. However, the DADM recognises that lower-impact algorithms may not need direct human involvement.
- Like the Charter, the DADM requires peer review. However, the DADM goes further by specifying which levels of impact require peer review and which experts are qualified to conduct it, with required expertise increasing as potential impact increases.

On the other hand, New Zealand's Charter was generally seen as an effort to promote public trust rather than as an impact control framework for the GC. Therefore, New Zealand's approach encompassed a broader range of normative commitments:

- Active engagement with communities impacted by the algorithm's use, which the Charter explicitly calls for. This principle is present only in the GC framework when the need arises in an impact assessment.
- Impacted communities may also have Treaty interests, with the Charter calling for embedding an indigenous perspective in the development and use of algorithms. Indigenous-considerate development is not required in the GC framework.

In the age of generative AI, whose capacity to generate "unknown unknowns" challenges established risk methodologies that rely on knowing unknowns and controlling them, AI governance can no longer rely solely on this method. It also needs a normative framework that offers guiding principles to shape the acceptable use of AI systems: one that aims to maximise public benefit and uphold public trust, rather than just minimise the residual risk posed by AI systems to citizens. Given New Zealand's relatively low trust in AI and declining trust in government, a trust-enhancing approach to AI system delivery is even more critical.

However, normative frameworks cannot replace organisational methodologies—they must support and inform them. The practical challenge of implementing the Charter's norms was highlighted in the Year 1 Review of the Algorithm Charter by Taylor Fry (2021). Nonetheless, this review confirmed the Charter's overall intent and approach.

The normative commitment to enhance public trust embodied by the Algorithm Charter is more important than ever in the age of generative AI, declining public trust in government, and low public trust in AI relative to global peers.

2.1. Action: Fold Algorithm Charter improvement work into the Public Service AI work programme

Operationalising Year 1 Review findings has been challenging, with recommendations only partly implemented. Table 5 monitors the progress of the considerations identified in the review, excluding those that simply confirm the current direction. Of my seven groups of considerations, only two have been completed. Two show progress mainly driven by efforts outside of the system lead agencies. Three show no publicly available evidence of progress.

Done?	**Grouping of actionable consideration(s) in Taylor Fry (2021)**	**Progress as of January 2025**
☑️	Maintain a risk-based approach with supplementary tools for risk assessment and further clarifying guidance that acknowledges the value of algorithms.	Completed in December 2023 with the release of the AIA toolkit, with the Algorithm Threshold Assessment and AIA user guide respectively filling the gaps.
☑️	Facilitate a community of practice to share use cases captured in the Charter.	BAU – meets quarterly but no public updates since September 2023.
➡️	Consider working with Māori data experts to develop more detailed guidance, best practice on partnering with Māori, clarify working with existing consultation groups, sharing knowledge with other agencies given limited expert capacity in this area.	Te Kāhui Raraunga has independently developed the Māori AI Governance model. In the same report, **they call for an urgent overhaul to the Algorithm Charter** to align with their model.
➡️	Detailed technical resources for measuring bias and ensuring human oversight of algorithm outputs.	For traditional machine learning, MSD's Model Development Lifecycle is sufficiently detailed, and the AIA toolkit mentions this. No equivalent technical resources relevant to generative AI exists, but the GCDO is working on an AI assurance model.
❌	Investigate novel forms of citizen participation and measuring public trust and confidence in algorithm use.	The Digital Council report investigating this for automated decision-making was shelved and the Digital Council was dissolved. DPMC's PCET may be relevant.
❌	Consider the creation of an oversight body for the Charter.	No public knowledge of consideration. GCDO has appointed an AI Expert *Advisory* Panel, but it does not have any oversight powers.
❌	Develop an annually updated algorithm register.	The GCDO has committed to a central AI register by 2027.

Table 5: Actionable considerations identified in the year 1 review of the Algorithm Charter. "Actionable" excludes passive considerations that recommend maintaining current direction. From Taylor Fry, 2021. Algorithm Charter for Aotearoa New Zealand Year 1 Review. https://www.data.govt.nz/assets/data-ethics/algorithm/Algorithm-Charter-Year-1-Review-FINAL.pdf

Over the same period, the GC has adopted a relatively proactive approach to maintenance, completing four reviews of the DADM. Two of these reviews were conducted in the generative AI era following the release of ChatGPT. Like New Zealand, the GC has issued specific guidance on the use of generative AI. However, unlike New Zealand's guidance, their GenAI guidance clearly separates it from existing algorithmic guidance. The GC's guidance also offers advice on the

most suitable contexts for applying these techniques. Section 3 of this report evaluates the progress of the New Zealand Government's generative AI guidance.

While Canada has kept its government automated decision-making guidance up-to-date and integrated with new generative AI guidance, New Zealand's algorithmic guidance has not kept up.

2.2. Action: Conduct an evaluation of the AIA toolkit with a view to folding it into the NZAIM

The AIA toolkit represents the most significant advancement in all-Government AI policy since the Algorithm Charter, turning high-level commitments into tangible actions. Its first tool, the algorithm threshold assessment (ATA), addresses a major criticism by clarifying the scope of systems covered by the Charter. The ATA's "material impact" test aligns with, but is broader than, the Ombudsman's interpretation of "personal capacity" in the OIA.

However, no review of the two-year-old AIA toolkit has been announced, despite technological advances, public service AI policy development, and public service AI adoption. Its author, Frith Tweedie, recently joined the GCDO's AI Expert Advisory Panel, indicating better alignment between GCDO and GCDS work programs.

Issues that a review of the AIA toolkit may explore include:

- Evaluating the effectiveness of the toolkit among end users, including the communication and engagement around this tool. Currently, no published AIAs exist for any agency. Agencies that have implemented AI tools have only carried out PIAs.
- Incorporating the Māori AI Governance model throughout the toolkit, not just through *Partnership with Māori*. For example, Question 6.3 (Storage) addresses jurisdictional issues but should also consider iwi Māori interests in onshoring data to bolster control. No Māori data experts were acknowledged for providing feedback on the original AIA.
- How it interacts with technical guidance instruments. For example, the use guide relies on MSD's Model Development Lifecycle for technical guidance, such as defining a model's accuracy. However, the MDL has not been updated since October 2021. Thus, it lacks technical guidance on newer techniques such as GenAI. A new NZAIM should serve as the definitive technical instrument, under which an impact assessment process sits alongside other universal processes, such as an AI-specific certification and accreditation process.
- How it interacts with strategic instruments. Given that the privacy impact assessment toolkit organises its risks around the information privacy principles, risks identified by the AIA may be similarly organised under the PSAIF principles.

The AIA toolkit is the strongest AI guidance instrument the Government currently maintains, but it must be evaluated to understand why no agencies have published a worked example in its two years of existence.

2.3. Action: Promote publication of algorithm threshold and impact assessments to data.govt.nz, creating a register that establishes implicit accountability

The Taylor Fry review found that public reporting of each agency's use of algorithms, as recommended under the Transparency commitment of the Charter, was "fragmented and incomplete" in 2021. As of December 2025, a site-specific search for "algorithm" on each Charter signatory's website yielded only one example of a self-reported agency-wide stocktake: the Ministry of Justice. agencies do make certain algorithm and AI assessments available, often in response to OIA requests regarding the Charter. However, consolidated information should already exist for agencies that responded to government-wide stocktakes conducted by system lead agencies for the 2018 algorithm assessment report[10] and the 2024 cross-agency survey on AI[11]. Consolidated information on high-impact algorithms (i.e. algorithms that "must" follow the Charter) should already exist within agencies' enterprise risk registers.

Proactive transparency about AI use acts as a low-effort mechanism of implicit accountability when legal enforcement is absent. An all-of-government register has been recommended since Gavaghan et al. (2019) and the Taylor Fry evaluation. This register can help spread best practices in tackling common challenges by sharing impact controls used in such cases. Such a register also creates a soft compliance dynamic, with agencies that do not report signalling lower governance maturity, thereby encouraging remediation as a form of reputational risk management. Comprehensive reporting on the use of algorithms and AI can also build public trust by clearly indicating where these systems are used and where they are not. Importantly, such a register can be developed without revealing sensitive details. At a minimum, a register should disclose:

- Basic context regarding when, where, and for what purpose the algorithm is used.
- Answers to the questions in the algorithm threshold assessment.
- If ATA's threshold is met, specify when and who (job title or qualification sufficient) conducted the latest peer review to evaluate potential unintended consequences and how they responded to this information, or justify why a review has not been recently performed (e.g. static legislation and environment minimise risk of concept drift).

In this regime, information may still be justifiably withheld in accordance with the OIA. However, the public can be assured that a separate expert found the algorithm to be, at a minimum, technically and legally compliant. A peer review can offer assurance regarding the lawfulness of decision-making subject to judicial review, in accordance with its establishing legislation and the broader legislative framework previously mentioned. A peer review may also examine alignment with the Charter's non-mandatory commitments. The AIA toolkit already provides a sufficiently comprehensive framework to organise such a peer review.

---

[10] *StatsNZ, 2018. Algorithm assessment report.* *https://data.govt.nz/toolkit/data-ethics/government-algorithm-transparency-and-accountability/algorithm-assessment-report*

[11] *Department of Internal Affairs, October 2024. Proactive release of material relating to Artificial Intelligence in the month of July.* *https://www.dia.govt.nz/diawebsite.nsf/Files/Proactive-Releases-2024-25/$file/Proactive-release-of-material-relating-to-Artificial-Intelligence-in-the-month-of-July.pdf*

In Canada, the DADM mandates transparency by requiring the publication of algorithmic impact assessments on the federal Open Government Portal. This website uses CKAN, which also powers New Zealand's data.govt.nz dataset catalogue, maintained by the GCDO. It is straightforward to add an "Algorithm" or "AIA" tag within data.govt.nz to emulate Canada's ADM register. Agencies already have procedures and APIs for uploading content to this portal.

Data.govt.nz can already be used to create a minimum viable self-reported register of algorithm and AI use across government with little additional work, given agencies should already know the necessary information.

2.4. Action: Refactor the Algorithm Charter as an extension of DPUP: normative values for AI system delivery to follow

Both the Charter and the DPUP fulfil closely related functions. Each articulates expectations for the use of data and algorithms that extend beyond existing obligations, with the aim of promoting public trust through their respectful and safe use. However, each fall short in ways that the other compensates for. Taylor Fry (2021) observed that the Charter lacks clarity around the practical implementation of its commitments to partnership and community engagement. The subsequent AIA toolkit now points to DPUP for guidance in these areas. On the other hand, DPUP lacks the formal mandate of other standards, or, in the case of the Charter, publicly recorded voluntary commitment by agencies. DPUP has been employed by agencies implementing AI outside of its original remit in the social sector, such as ACC.[12]

In recognising their similarity, compliance can be streamlined by refactoring[13] the normative Charter commitments as a function of DPUP. The Charter commitments of partnership and people overlap with the Manaakitanga principle of DPUP, which goes into greater detail around how to engage communities with a view to upholding their dignity and mana. The Charter's call to "deliver clear public benefit" is given greater emphasis under the He Tāngata principle in DPUP, which promotes a focus on improving people's lives. The He Tāngata principle provides a clearer constraint on what AI systems should optimise for, prioritising life outcomes over other definitions of public benefit, such as naïve fiscal optimisation.

A streamlined, enforceable DPUP will emphasise the fundamental, foundational nature of GCDS guidance centred on maintaining public trust in the use of information. Trustworthy use of any information technology does not happen without the trustworthy use of the information that communities entrust to agencies. Conversely, comprehensively trustworthy use of this information begets the trustworthy use of any novel information technology, provided that suitable engagement and evaluation mechanisms are already in place. Trustworthiness can be demonstrated by technical practitioners by ensuring safety, effectiveness, and accountability. However, systems can demonstrate all three without being trustworthy if their use is fundamentally inappropriate. Appropriateness is DPUP's reason for being: by describing "what

[12] *ACC, March 2025. ACC Privacy Impact Assessment (PIA) - Agent Copilot. https://www.acc.co.nz/assets/corporate-documents/Privacy-Impact-Assessment-Agent-Copilot.pdf*

[13] *Reorganising text (typically software code) to improve readability and reusability without affecting its function.*

doing the right thing looks like". Technical practitioners alone rarely possess the capability or institutional authority to set such cultural expectations for respectful data practice and to sustain active, meaningful community engagement. Setting cultural expectations requires an explicit whole-of-agency commitment and support, sponsored by a single organisational leader to steward best practice – regardless of whether data is used by algorithms, AI, or humans directly. A unified all-of-government policy provides a coherent structure to support agencies in developing and sustaining such a trustworthy data culture.

A consolidation of GCDS policy also provides an opportunity to consider other pending work:

- How to promote Ngā Tikanga Paihere as general advice for the wider data system, not just projects within the Integrated Data Infrastructure: as a function of DPUP+.
- Consolidating practical advice on Māori engagement and citizen participation, building on existing all-of-government initiatives like the DPMC's Policy Project[14], and previous technology-specific research by Toi Āria[15] and Te Kāhui Raraunga[16].
- How compliance with GCDS guidance is assessed within implementing agencies, in lieu of legislative, Cabinet or voluntary enforcement. Another potential mechanism is a standardised data maturity assessment, through which agencies are evaluated, directly associating a trustworthy, respectful data culture with high performance. It is unclear how or if the GCDS currently assesses organisational data maturity.

After streamlining normative commitments, what remains of the Algorithm Charter comprises operational obligations that are already mandated by legislation and standards and thus are best clarified in the NZAIM. Transparency around algorithmic decision-making is largely compulsory under the OIA. Failing to ensure and regularly review the fitness of data and algorithms constitutes poor professional practice and exposes agencies to material risk. Privacy and the human right to non-discrimination (through bias mitigation) are safeguarded by existing laws. Accordingly, these residual commitments are largely addressable through technical controls rather than requiring standalone normative commitments.

The GCDS maintains a range of normative best practice on specific systems (algorithms, IDI) which are largely applicable to any data use. All guidance promoting a trustworthy, respectful data culture can be consolidated into DPUP.

[14] *Department of Prime Minister and Cabinet, October 2023. Policy Community Engagement Tool. https://www.dpmc.govt.nz/publications/policy-community-engagement-tool*

[15] *Digital Council for Aotearoa New Zealand, 2020. Towards trustworthy and trusted automated decision-making in Aotearoa. https://www.toiaria.org/our-projects/towards-trustworthybrand-trusted-automated-decision-brmaking-aotearoa/*

[16] *As outlined in Table 4.*

3. The GCDO's PSAIF and RAIG are not currently a unified guidance framework

The Public Service AI Framework (PSAIF) is a strategy, not an implementation framework. It offers an achievable vision, grounds it in the OECD AI Principles and our unique legal context, and identifies a work programme to realise the PSAIF's desired outcome. Viewed as a strategy, the PSAIF is fit for purpose as it provides strategic direction and defines what good looks like. While it helps agencies devise their own AI strategies, it does not specify how agencies should execute this strategy. For example, the PSAIF correctly identifies the OIA as a relevant law that applies to AI. However, the PSAIF does not explain how the OIA places significant constraints on AI usage, as in Box 1.1.1.

Instead, the Responsible AI Guidance (RAIG) – a suite of guidance with the first in the series focusing on Public Service GenAI (RAIG-PSG) – outlines the 'how'. It shows potential as a comprehensive, unified guidance framework. Its GenAI guidance provides useful, novel advice for that technology and references existing guidance such as the NZISM and GWS. However, RAIG-PSG does not sufficiently incorporate best practice from the earlier Algorithm Charter or the later RAIG for businesses. Furthermore, the RAIG-PSG mentions but does not organise its guidance around the principles of the PSAIF.

Section 2.4 outlined how a redeveloped GCDS policy should interact with the GCDO framework. GCDS policy should uplift organisational data culture, and GCDO guidance should acknowledge how AI developers should employ existing mechanisms recommended by GCDS policy. This section outlines how to harmonise the frameworks that the GCDO is responsible for.

3.1. Action: Align guidance for businesses and the public service

Some of the best guidance for responsible AI delivery comes from an agency that has no system leadership over AI in the public service – the Ministry for Business, Innovation and Employment (MBIE). As the leader of microeconomic policy, MBIE has been tasked by Cabinet with helping businesses use AI responsibly. This work culminated in the National AI Strategy and the Responsible AI Guidance for Businesses (RAIG-B), released in July 2025. This separation of roles has not resulted in markedly different guidance despite the contrasting risk profiles and incentives in the public and private sectors. On the contrary, RAIG-B offers useful technical guidance applicable to the public sector and could be adopted more broadly in a comprehensive NZAIM. Some points emphasised in RAIG-B are not as clear in the current RAIG-PSG, including:

- An emphasis on ensuring high-quality, fit-for-purpose training data, and framing bias and unfairness as originating from poor quality data – which leads to poorly performing AI. RAIG-PSG considers this an independent concept that is only mitigated through process controls. RAIG-B correctly highlights that technical controls are equally effective, providing useful examples such as "a facial recognition model to be used in New Zealand would likely be more accurate and effective if trained on images representative of the New Zealand population".
- Encouraging consideration of the legality and ethics of certain data collection and use. While RAIG-PSG recognises agencies' Privacy Act obligations, it does not acknowledge other sorts of legally and ethically contentious data collection identified in RAIG-B, such as

using copyrighted work without permission and being cautious with web scraping. RAIG-B emphasises that options are available for obtaining ethically trained AI systems. Its advice on Māori data is more comprehensive and affirming of Māori sovereignty than Crown guidance, despite only the Crown having formal commitments to iwi Māori.

Responsible AI guidance for businesses also provides useful technical advice for government agencies. This advice should be integrated into the NZAIM.

3.2. Action: Reorganise all guidance around the PSAIF to provide clarity on how agencies meet the expectations of the strategy

While the PSAIF and RAIG-PSG are fit for purpose as independent artefacts, reading them in conjunction can be difficult, as they are not structured similarly. The two were released simultaneously from the same agency, so it is unclear why the RAIG-PSG has opted to use the OECD AI Principles, rather than the principles of the PSAIF, which have been "inform[ed]" by the OECD principles but ostensibly modified for the New Zealand context.

The PSAIF prompts the consideration of legal and regulatory instruments, but the RAIG-PSG does not acknowledge important laws that enforce its guidance, such as the transparency obligations arising from the OIA, the anti-discrimination obligations from the Human Rights Act, and only references the Privacy Act once. The RAIG-PSG does not mention the Treaty of Waitangi or relevant Waitangi Tribunal findings, which the PSAIF identifies as significant constitutional context. The RAIG-PSG does not mention "social licence" which the PSAIF identifies as one of its six pillars.

As discussed before, the AIA toolkit can be reorganised around the PSAIF principles. This approach aligns with the PIA toolkit, organised around the information privacy principles.

As the suite of responsible AI guidance expands, greater structural consistency across guidance artefacts can improve usability and coherence of the GCDO's overall strategy and guidance ecosystem. Clearly aligning subordinate guidance with the PSAIF can enhance this framework's operational effectiveness by translating its strategic intent and vision into clear actions for agencies. Structural alignment may also further enable agencies to responsibly innovate with novel forms of AI (such as the emerging techniques discussed in Section 4.3), by establishing precedents through which the PSAIF can be given practical effect, rather than agencies relying on prescriptive guidance, each with its own unique structure.

The PSAIF provides a flexible framework to organise AI guidance. The RAIGs and the AIA toolkit should be organised around the principles of the PSAIF.

4. Like the NZISM, the NZAIM should delineate the different types of algorithm and AI techniques

The NZISM defines common cybersecurity procedures across information technology systems. It separately identifies specific controls for particular types of systems, such as software, email systems, networks, and gateways. The NZAIM should adopt this unified yet specific approach.

Guidance should distinguish between techniques only as precisely as is practical to minimise the compliance burden. A useful organising principle is the evaluation paradigm under which a system is assessed. The methods an agency uses to ensure system performance largely determine both the operational guidance relevant to development and the monitoring and governance arrangements needed on an ongoing basis. Importantly, these methods tend to be consistent within a given evaluation paradigm, enabling common best-practice guidance, assurance and governance approaches to be applied across a diverse range of techniques.

For example, ACC's use of algorithms and AI spans from the most basic to the most advanced. When a claim is sent to ACC, **business rules** search the claim's free-text accident description[17] for specific terms to code information into structured data. Such algorithms are simple to develop and evaluate using traditional testing. However, they still require continuous validation. Here, manual validation is required to identify new terms not covered by existing rules, such as "collided with Flamingo [scooter]" or "Tesla autopilot crashed into a pole". ACC can substitute more complex algorithms, such as fuzzy matching, but the evaluation methods remain the same.

If a claim is successfully coded by these rules, ACC immediately processes the claim using a **supervised model** (logistic regression) to determine auto acceptance[17] or hold for manual assessment. Here, the rules are now automatically generated by optimising a discrete output (accept or hold) based on past data characteristics. These rules are also straightforward to assess for accuracy: by applying them to data it has not seen before to measure overall accuracy, and over protected attributes like sex and age to monitor fairness. ACC can substitute more complex models in this system, but the evaluation methods remain the same.

Any accepted claimant with more complex needs may need to contact ACC over the phone. Their calls will likely be transcribed[18] by a **generative AI model** with millions to billions of times more complexity than the previously mentioned algorithms. The open-ended nature of generative AI makes it harder to empirically measure performance in the context in which it is deployed, as is the case for non-generative AI. Instead, evaluation relies on expert and user acceptance of sample answers for a defined use case. ACC can increase the system's flexibility and utility, such as augmenting the call transcript with its knowledge base, or changing to a different model architecture with improved language translation, but the evaluation methods remain the same. To remain flexible to technological developments, guidance should not descend into finer technical distinctions beyond the evaluation method.

---

[17] *ACC, August 2018. Statistical models to improve ACC claims approval and registration process. https://s3.ap-southeast-2.amazonaws.com/nzdoctor.files/production/public/2018-08/claims-approval-technical-summary_0.pdf*
[18] *ACC, March 2025. ACC Privacy Impact Assessment (PIA) - Agent Copilot. https://www.acc.co.nz/assets/corporate-documents/Privacy-Impact-Assessment-Agent-Copilot.pdf*

4.1. Action: Adopt a taxonomy that recognises commonalities between algorithms, traditional and generative AI

Given these common controls across algorithms and AI, I recommend that AI system guidance cover traditional algorithms. This distinction is not evident in existing AI-specific guidance. This distinction is not solely a matter of academic precision, but a practical safeguard. Traditional automations outside the modern understanding of AI still have the same potential for beneficence (e.g. freeing up workers from menial, repetitive tasks, standardising and de-biasing processes) and maleficence (e.g. Robodebt[19], Dutch childcare benefits scandal[20]) as AI. Promoting this understanding ensures that governance and monitoring efforts appropriately cover all systems with the potential for material impact, not only those labelled as AI. Furthermore, the same legal obligations apply to any impactful decision-making system, regardless of whether it is a traditional algorithm or an AI system.

A similar risk arises when conflating AI with generative AI. AI systems have long been deployed across government using traditional predictive techniques, as outlined in Section 5. Focusing efforts on newer generative AI systems, currently limited to low-impact administrative uses, while ignoring the monitoring of established predictive AI systems, risks diverting oversight away from systems that already make consequential decisions about users of government services.

Figure 2 illustrates a categorisation that recognises the similarities and differences among these evaluation paradigms, visualising the nesting and overlaps among the categories. Below, I propose definitions for key categories of algorithms and AI, and discuss:

- what other categories of techniques fall into that category (**subsets**)
- what other categories share techniques in that category (**intersections**)
- **examples** of techniques in that category
- methods for **evaluation** of model suitability, both during design and operation
- **challenges** universally associated with designing the systems included within that category. Subcategories inherit the challenges of their parent category. For example, the issues identified for algorithms (automation bias, auditability, monitoring and evaluation) extend to all techniques mentioned in this paper.

---

[19] *Australian federal government data matching algorithm between one's declared income to the social service agency and actual income known to the tax agency to calculate welfare overpayments. This algorithm made invalid assumptions about income, and its validity was never tested. This scheme triggered parliamentary inquiries and a Royal Commission report, which became a political issue in the election where the responsible government lost.*

[20] *Dutch government risk-scoring algorithm to determine the risk of childcare benefit fraud. Rules were manually developed by humans, factoring in protected attributes like dual nationality. Rules and outputs were not made available to flagged households, and decisions could not be contested. Staff succumbed to automation bias, treating the outputs as evidence of fraud rather than a signal to investigate actual evidence of fraud. Consequently, parliamentary opposition moved a motion of no confidence in the government, forcing the serving government to resign.*

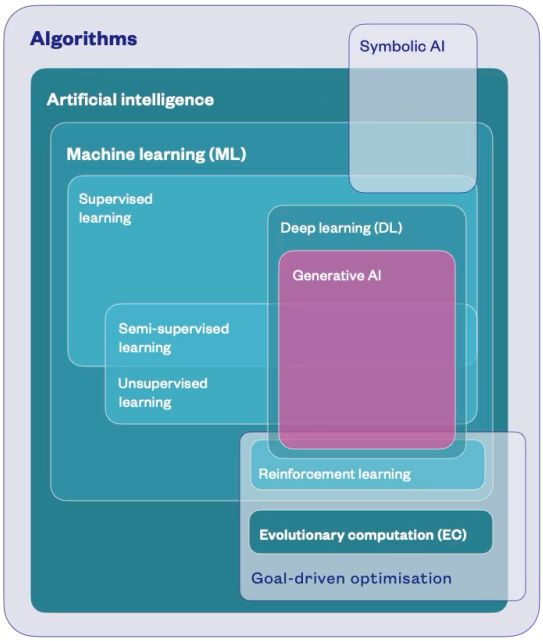
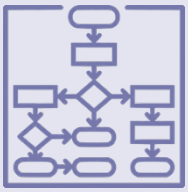


***Figure 2**: Overview of relevant categories for algorithmic and AI techniques with distinct evaluation and interpretation considerations. Algorithms form the superset of all analytical techniques discussed in this paper, within which AI constitutes a subset.*

Two of the most relevant paradigms in AI are shown. The first is machine learning (ML), the main set of the most used AI techniques. The second is evolutionary computation (EC), a distinct type of AI used for optimisation and simulation.

Relevant types of ML are also shown: supervised learning, unsupervised learning, reinforcement learning. Two other categories employ a mix of those three techniques in an advanced, novel manner: deep learning (which itself is not an evaluation paradigm but a type of model), and generative AI. Goal-driven optimisation is a distinct evaluation paradigm.



Box 4.1.1. Algorithms

Methodical set of instructions that can be executed by a machine. Algorithms are typically authored by humans but increasingly can be written by generative AI.

Subsets (non-exhaustive): goal-driven optimisation, artificial intelligence

Examples: any logic that runs on machines, from a simple business rule to categorise an email, to a generative AI model powering advanced features in existing software.

Evaluation: varies by technique. At a high level, algorithms are validated using traditional software testing frameworks. Test frameworks typically distinguish between verifying the outputs of the system and validating the outcomes the system sought to achieve. Tests are typically manually written but increasingly can be written by advanced AI.

Challenges:

- *wide definition* makes it hard to track for governance purposes
- *lack of continuous monitoring* risks performance degradation as the operating environment changes from when the rules were developed (concept drift) or encounters new data it cannot properly handle (data drift)
- *lack of documentation* may deny individuals their official information rights to rules that lead to, or the reasons for, decisions or recommendations
- *poor auditability* without an audit trail required for all public records, including those generated by an algorithm
- *automation bias* may result in uncritical human deference to automatic predictions
- *appeals, judicial review and Ombudsman investigations* may be taken against a decision, recommendation or action performed by an algorithm, ensuring that such acts were made lawfully. A degree of reversibility may be required.

Box 4.1.2. Predesigned / handcrafted algorithms

Algorithms consisting of predefined logic, rules or 'symbols' (as in the traditional category of symbolic AI) for reaching an output. These algorithms are typically manually designed by humans but may be designed by large language models through linguistic likelihood, rather than mathematical derivation. This category excludes data-derived systems that may itself consist of rules or symbols, such as decision tree classifiers.

Examples: expert systems, business rules, predefined decision trees

Evaluation: as for algorithms

Challenges: relative simplicity and explainability may come at the expense of accuracy.

Box 4.1.3. Goal-driven optimisation (GDO)

Algorithms that search for the best solution to a defined problem, which becomes exponentially difficult when attempted exhaustively. They typically use simulations to model a solution's interaction with its environment. The "best solution" may be the end intervention itself, such as a set of school bus routes that maximise coverage and minimise resource use. It may also be the best strategy for how actors behave and react to an intervention.

Actors[21] optimise their behaviour based on their own motivation (e.g. economic gain, travel time minimisation, personal health) reaching equilibrium against others in the simulation.

Subsets: reinforcement learning, evolutionary computation.

Examples: navigation application planning, supply chain optimisation, school bus route generation (Ministry of Education's School Route Transport Optimiser), transport system modelling (Ministry of Transport's Monty), large population models (PHF Science's ALMA)

Evaluation: the output is necessarily the 'best' given the problem and search constraints, but further validation is required to ensure the desired outcome was met (cf. impact evaluation in policy design)

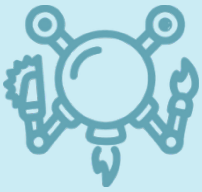
[Challenges continues next page]

[21] *I avoid using the term 'agents' and defining 'agentic AI' as a category. Agency describes a philosophical capacity to act with intention, rather than a description of technical evaluation. This taxonomy instead distinguishes between orchestrating LLMs and GDO models. LLM-orchestrated 'AI agents' (such as Microsoft Copilot Researcher or Github Copilot) only reason the next likely action based on probabilistic likelihood, as explained in Box 4.1.9. Even though they can perform actions, orchestrating LLMs lack the independent intent to measure and satisfy a goal beyond prompt completion. In contrast, GDO models learn how to act within an environment, real or virtual, guided by a defined mathematical objective.*

For example, if an 'agent' is tasked with devising an individual's unique injury rehabilitation plan, an orchestrating LLM would generate a plan mimicking prior examples with similar circumstances. A GDO model would simulate the patient's unique biomedical, psychological and social risk factors and calculate how specific treatments and programmes optimise the probability of returning to independence. Either approach is valuable in different situations: LLM orchestration provides a user-friendly natural language interface that draws on historical patterns, while GDO provides systemic mathematical fidelity that models the actual effect of a plan.

Challenges:

- *defining the true problem*, constraints and objectives are difficult, requires translating end-user requirements to measurable equations, which can often be incomplete, framed poorly or from a deficit basis, can change over time
- *representativeness* of simulated population and behaviour to actual population
- *computationally expensive*, often using AI methods to more quickly find a sufficiently (not always the most) effective solution



Box 4.1.4. Artificial intelligence (AI)

Machine-based systems that infer (based on implicit or explicit objectives) from the input it receives how to generate outputs (predictions, content, recommendations, decisions, actions).

Subsets: machine learning, evolutionary computation

Examples: from basic linear regressions or probability models that predict numbers or outcomes; to advanced deep neural networks that can generate complex nuanced text and images, or guide decision-making of simulated actors.

Evaluation / challenges: varies by technique

Box 4.1.5. Machine learning (ML)

AI models that are automatically developed using existing observations, understanding and approximating how that data contributes to an output.

Subsets: supervised learning, unsupervised learning, reinforcement learning, deep learning, generative AI

Evaluation: self-evaluates its **output** at each iteration of model development using the objective it was given, but further manual evaluation is typically required to assess whether the objective has produced the right **outcome** (cf. impact evaluation in policy design)

Challenges:

- *data accuracy* – predictions are only as reliable as the underlying data, risk of systemic error (e.g. police non-completion of callout assessment correlated with victim ethnicity), human error (e.g. staff assumes data points instead of asking the subject), and faulty assumptions (e.g. using outputs as proxy for outcomes).
- *data robustness* – data must be representative of the population and operating environment which a model will be applied (e.g. oversampling under-represented groups, discarding stale data). *[Challenges continue next page]*

- *confounding variables* – hidden factors influencing variables may result in misleading conclusions, such as home environment stability confounding the relationship between school attendance and educational achievement.
- *fairness and equity* – predictions often reflect the biases within the data and the wider system where the data originates from; as it is difficult to satisfy all the different definitions of fairness, a normative decision is required for what kind of equity is desired and optimised for.
- *accountability* – as no human was directly responsible in generating the decision-making process making, responsibility and liability is less clear.

Box 4.1.6. Supervised learning

ML models that infer how to generate outputs based on the relationship of previously observed inputs with an associated output.

Subsets: self-supervised learning (associated output comes from the input data, typically used in generating large complex sequences like languages, images)

Examples: classifiers (decision trees, gradient boosting: e.g. StatsNZ International Migration predictions, logistic regression: e.g. ACC Probability of Accept / RoC*RoI), regressions (linear regression)

Evaluation: measured based on how well the model performs on past but previously unseen examples of data.

Challenges:

- *generalisation* – careful testing methodology required to ensure patterns are learnt (i.e. not just examples memorised) that can be accurately applied outside of training and for extreme edge cases.
- *validity of targets* – assumes target outputs are correct and consistent, such as adjusting for environmental or legislative changes since the data was captured
- *suitability of targets* – critically assessing whether the chosen predicted output (e.g. risk of reimprisonment and reconviction) aligns with broader desired outcomes (e.g. releasing an offender will not result in future societal adversities), where the output variable may introduce confounding (e.g. reimprisonment may be less likely in less policed areas).

Box 4.1.7. Unsupervised learning

ML models that infer how to generate outputs from the input it receives without explicitly knowing what outputs to generate (from past examples), based only on underlying patterns in the input data. [22]

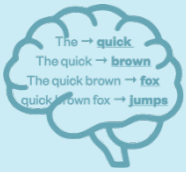
Subsets: semi-supervised learning (a small, labelled dataset provides the starting point for unsupervised learning on a larger unlabelled dataset)

Examples: anomaly detection (e.g. fraud and abuse detection), clustering (e.g. cohort detection for targeted policy interventions)

Evaluation: measured by how well the model captures structure and patterns in the data (e.g. separation of clusters, reconstruction quality). Developing robust objective metrics is more difficult during model development because, unlike supervised learning, there is no known ground truth. A semi-supervised approach (labelling a small dataset) may be used for known concepts like fraud and abuse.

Challenges:

- *fairness and bias* – despite seeming inherently unbiased by learning from underlying patterns, it is still susceptible to learning systemic biases embedded in the input data, or statistical biases from a lack of sufficient representation.
- *imbalanced data* – patterns of interest (e.g. fraud and abuse) may be overshadowed by patterns of normal behaviour, making detection or pre-definition difficult.



Box 4.1.8. Self-supervised learning

ML models that infer how to generate outputs through reproducing input data rather than relying on an independently provided target. As the target output comes from the input data itself, this ML paradigm combines the verifiability of supervised learning (withheld data can be used as ground truth) and the scalability of unsupervised learning (no manual labelling and eliminates bias from proxy labels).

Examples: LLM pre-training, large population models (e.g. PHF Science's ALMA)

Evaluation: measured by how well the model recreates previously unseen input data.

Challenges: fairness and bias as with unsupervised learning

[22] *This framework differentiates between supervised and unsupervised learning models to recognise the findings of a review by the Australian National Audit Office (ANAO) into the Australian Tax Office's (ATO) use of AI. The ATO's use of AI comprised 71% unsupervised learning models, by virtue of their role in identifying non-compliance where patterns of concern may be hard to detect, poorly defined, or emerge from novel, unanticipated behaviour. Presumably due to the challenges of measuring these models' performance, the ATO did not develop ways to measure any of the models' performance. Adaptive identification of non-compliance will be desirable for – if not already used by – many government agencies here. Therefore, standard guidance should recognise the increased risk around the development of unsupervised non-compliance models.*

Box 4.1.9. Generative AI (GenAI)

DL models that create new data (text, images, video, music, speech) by learning and mimicking the underlying patterns within existing data.

Examples: large language models (e.g. GPT-5, Claude Sonnet), orchestrating LLMs (e.g. Microsoft Copilot Analyst, Github Copilot), diffusion models (e.g. Midjourney, OpenAI Sora)

Evaluation: as with all ML techniques, but output evaluation is much more difficult due to the near-infinite and unpredictable nature of its output. The tasks these models have been trained for (e.g. generally predicting the next likely word in a sequence) may be different from the tasks they are used for in practice (e.g. analysing policy intervention options, performing calculations on tables pasted into chat). Outcome evaluation is therefore more important and is typically done manually (and is always manual when using off-the-shelf models that the deployer does not train further).

Challenges:

- *hallucination* – production of plausible but false content, exacerbated by the potential for deceptive confidence: delivering false information as authoritatively and fluently as faithful information; and the subtlety of potential errors given its optimisation for linguistic likelihood rather than factuality.
- *homogenisation* – generating overly uniform content leading to the marginalisation of underrepresented ideas and identities by regressing to the mean.
- *user hijacking* – manipulating the system with inputs that result in undesirable behaviour (e.g. prompt injection, data poisoning)
- *supply chain provenance* – lack of transparency around the reliability of constituent components of a GenAI system (e.g. are datasets and pre-trained models accurate, representative, and obtained legally and ethically).
    - Provenance is especially important if outputs may derive from:
        - protected mātauranga Māori (e.g. language, art, knowledge) as permission must be sought from the authors or kaitiaki of such materials
        - tapu materials which should never be used to create new outputs
- *generation of dangerous content* – including dangerous weapons; dangerous, violent or hateful content; misinformation and disinformation; offensive cyber-attacks; obscene harmful imagery
- *unauthorised data integration or deanonymisation* – leakage of sensitive information from training data or inputs, both explicitly (e.g. publicly available social media profiles) and implicitly (e.g. semantic cues that imply personal attributes)
- *content attribution* – as part of the general challenge of auditability and recordkeeping.
- *environmental impact* – training models require vast computational resource (and thus energy or water for cooling); impact is less so for end use of pre-trained models but cumulative resource use can be substantial at scale.

Box 4.1.10. Reinforcement learning (RL)

AI models that learn how actors in an environment should effectively act (known as an actor's policy, e.g. writing the next best word, perform the next best economic transaction, wait or administer a medical treatment) rather than training on past examples.

Examples: large language models (RL from human feedback), macroeconomic models (e.g. Salesforce AI Economist), dynamic healthcare treatment personalised to patient characteristics and needs and real-time diagnostics

Challenges: *generalisability* – genuinely robust solutions are hard to distinguish over fortuitous solutions performing well under the conditions it encountered in a particular training run. This instability is worsened as each actor's actions directly shape the data it learns from, creating correlations that amplify noise.

Box 4.1.11. Evolutionary computation (EC)

AI models that mimic the trial-and-error, survival-of-the-fittest nature of biological evolution. Unlike traditional machine learning, which primarily optimises solutions mathematically, evolutionary computation randomly adjusts candidate solutions and retains a subset of the best ones. These methods can be especially useful when conventional ML reaches a dead end and cannot find a mathematical way to improve a solution, or when there is no clear mathematical formulation for how to improve it. Evolutionary computation maintains a pool of potentially effective candidates and evaluates them using externally defined measures of "fitness" that are not constrained by the internal structure of the solution.

Examples: actor behaviour in simulations (e.g. MATSim in Ministry of Transport's Monty traffic model: simulated transport system users determine their own daily travel plans based on historical data, 'compete' with other transport system users, everyone reaches an optimised equilibrium that balances everyone's travel objectives without pre-defining behaviour – beneficial for modelling infrastructure change), resource allocation and scheduling

Challenges: *genetic instability* – careful model design is often required to ensure beneficial traits are not lost to excessive variation, reducing the likelihood of settling on a truly effective solution.

4.2. Action: A new NZAIM should recognise the different technical challenges of different AI paradigms

Evidently, there are common challenges and considerations that apply to all algorithmic and AI techniques. For example, the risks of automation bias, concept drift, and output-outcome misalignment can be realised with any algorithm, regardless of the technique used. These risks are rooted in the broader system and environment where these algorithms operate, rather than in the algorithms themselves. However, these challenges often appear and are addressed

differently depending on the technique. For example, automation bias is relatively straightforward to tackle with a single prediction from supervised learning, compared to the active critical thinking needed to evaluate the open-ended output of generative AI.

These nuances require more detailed technical guidance than current one-size-fits-all approaches. This is not just about technical accuracy, but a user-focused need to reduce perceived compliance efforts, especially for non-generative AI, which already has decades of experience in trustworthy development. For example, while there are standard methods for assessing the performance and fairness of supervised machine learning, evaluating generative AI is far less straightforward and often depends on manual assessment in the specific context in which it is used. A new NZAIM must be designed accordingly, offering methodologies and controls appropriate for certain or all types of systems.

Like the NZISM, the NZAIM should firstly provide broadly applicable controls to all systems, then prescribe controls based on the risks unique to the evaluation paradigm used by an AI model.

4.3. Action: Consider the strategic use of simulations and goal-driven optimisation

Using this more precise taxonomy uncovers an opportunity for goal-driven optimisation to enhance the rigour of intervention appraisals and evaluations, thereby supporting more robust policy decisions, if not the rigour of intervention design itself. Instead of recreating past responses to interventions (as in supervised learning) or hypothesising policy impacts from emergent linguistic patterns (as in large language models), GDO models directly compute the effects of an intervention within a virtual environment.

This use case is not hypothetical, as agencies such as the Ministry of Transport use GDO (co-evolutionary algorithms) to simulate changes in traveller behaviour in response to a given transport system intervention. Vendors like PHF Science have developed new GDO models that lower the computational barriers to precisely and representatively simulate five million New Zealanders and their reactions to interventions.

Promoting effective and trustworthy AI use and decision-making with GDO requires system-wide coordination, as multiple agencies contribute data to these models and explore their use in isolation. Issues that a system lead agency can address include:

- Source data improvements: simulation actors are trained by replicating patterns from historical data and often rely on data sources that the interested agency does not collect, such as StatsNZ's Census data. The impending redevelopment of StatsNZ's census and survey programme presents a timely opportunity to consider how effective existing data sources are in simulations and GDO, and which new (or extensions to existing) datasets can have the greatest impact for a wide range of user agencies.
- Streamlining procurement: agencies currently commission bespoke GDO solutions tailored to their needs. The vendor market could be evaluated to see if vendor(s) can meet a wide range of agencies' requirements by offering domain-agnostic models that

provide flexibility across agencies' data and operational environments without creating bespoke model architectures for each use case.

- Encouraging use cases: existing communities of practice and networks can more intentionally promote this category of AI, which is currently limited to agencies responsible for explicit system design, such as the transport system. Any agency can reconceptualise their operations – or even policy remit – as an explicit system to be simulated, such as customer journey optimisation or regulatory impact management, with greater detail than traditional coarse appraisal models like Treasury's CBAx.

ML models predict to decide, generative models mimic to create, GDO models simulate to solve. GDO is already being used by agencies like MoT. System leadership can promote use and streamline development of GDO models.

5. Guidance must acknowledge different requirements across different use cases in which AI is applied

All-of-Government stocktakes in 2018 and 2024 have demonstrated the wide variety of ways that agencies already use or plan to use AI. The 2018 Algorithm Assessment Report clearly distinguished three different types of algorithms, consistent with my definition: operational algorithms, algorithms used for policy development and research, and business rules. This categorisation offers a useful contextual distinction between the aims of each kind of algorithm. Operational algorithms make or recommend complex decisions about individual users of government services. Policy algorithms make or recommend policy decisions that intervene at the aggregate system level. Business rules automate routine administrative processes without applying the same level of individual discretion as more advanced algorithms.

Six years later, the 2024 cross-agency survey of AI use cases adopted a different theming approach rather than the previous systematic taxonomy. While this theming was useful to illustrate the different uses of AI in the public sector, it is not scalable given its focus on specific objectives, such as "boosting productivity and efficiency" and "enhancing customer experience".

The 2024 survey also did not prompt agencies to consider different types of techniques; it only asked them to report on "AI use cases". This flaw in survey design has caused agencies to underreport the AI they use. For instance, only one AI use case was recorded by StatsNZ in the survey. However, traditional machine learning modelling has been used for official statistics, such as the International Migration provisional dataset, since 2018, generating highly publicised metrics for policy impact, like net migration. Use cases in ACC identified in Section 4, recognised in the 2018 survey, were not disclosed in the 2024 survey. Including them would have highlighted an AI use case that has clearly provided financial benefits to ACC, improved the experience for injured people navigating the system, and followed best practice. A clearer systematic definition of AI could enhance coverage of all kinds of AI use cases across government and highlight low residual risk, high-impact examples already integrated into agencies' day-to-day operations.

5.1. Action: A new taxonomy should differentiate between the different legal obligations associated with different use cases

A systematic taxonomy can ground AI standards and guidance on the specific considerations of the context in which AI is used. Below, I propose a new taxonomy that builds on the 2018 categorisation to accommodate technologies that have emerged since then.

- Operational algorithms now belong to the **"frontline"** category of techniques, which includes both algorithms and GenAI that directly and independently make decisions or recommendations regarding an individual that affect their rights or interests.
- Algorithms for policy development and research should also incorporate operational research using algorithms and AI, classified under the **"research"** category. This group encompasses both operational and policy research techniques that produce data to establish the evidence base for system-level decision-making.
- Most business rules and generative AI tools now fall into the **"administration"** category of techniques, which do not directly lead to a definitive decision but still have a significant impact. They may utilise personal data, subjecting it to the Privacy Act.

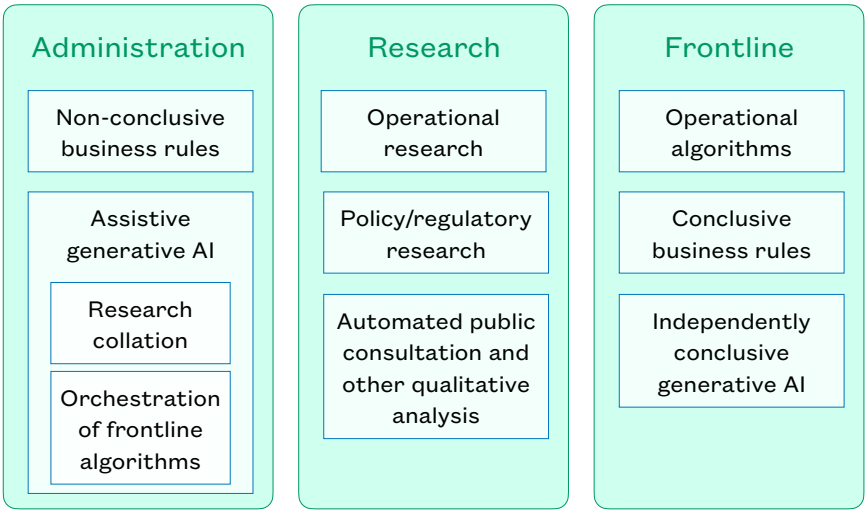


Figure 3: Taxonomy of use cases based on the relevant legal obligations within each category. Operational algorithms and any conclusive algorithm or GenAI are grouped into frontline. Evidence-generating algorithms and AI are grouped into research. Impactful but non-conclusive algorithms and GenAI are grouped into administration.

As in the previous section, I propose new definitions for these three categories, rooted in the relevant legal obligations associated with each, along with examples and the unique challenges each faces. Applications inherit challenges from the intersection they form with the technical taxonomy (e.g. all three contextual categories are susceptible to learning biases when developing supervised learning models).

Frontline AI directly makes decisions or recommendations around a user of government services, research AI generates evidence to inform system-level decisions, administrative AI support or automate the role of staff.

Box 5.1.1. Frontline

Algorithms and AI with its own set of rules that make decisions or recommendations about individuals (including corporations) that affect their rights or interests. Users have specific official information rights to the reasons behind these decisions. Fully automated decision-making and human-in-the-loop algorithms are included as there is no legal distinction between the two in the OIA. However, this category only includes GenAI systems that independently forms its own evaluative conclusions. GenAI systems that collate and summarise evidence to support a human decision maker is administrative.

Examples:

- **Front-line algorithms**: MSD automated decision-making (e.g. child support payments), Te Whatu Ora triage rules (e.g. Clinical Priority Assessment Criteria)
- **Supervised front-line models**: ACC Probability of Accept, border agency facial recognition
- **Unsupervised front-line models**: definitive fraud and abuse detection
- **Goal-driven front-line optimisation**: (hypothetical) dynamic clinical treatment policy personalised to patient characteristics and needs and real-time diagnostics, simulations appraising efficacy of a service provider's intervention
- **Front-line generative AI**: (hypothetical) risk profiling from trawling big data (call transcripts, client documents and assessments). **Should not be used in front-line without intrinsic, faithful explanations of its inferences.**

Challenges:

- **Accountability** – government decisions can be reviewed by **dedicated independent reviewers** (e.g. Benefit Review Committee for MSD benefit decisions). If such a channel is unavailable, the **Ombudsman** may also investigate decisions and actions, who may report back to the agency if they find a decision to be contrary to law, unreasonable, unjust, oppressive, improperly discriminatory, a mistake or wrong. Decisions and actions made under a specific legal power are also judicially reviewable, where the **High Court** assesses if actions were undertaken within the constraints of the law (e.g. algorithm's consistency with legislation). In any case, an oversight body will require a sufficient degree of explanation as to how a decision was reached.
- **Transparency** – any front-line decision made by an AI regarding an individual must be logically explainable linking the requestor's data ("material issues of fact") with every step of reasoning to the conclusion ("the reasons for the decision"). This requirement rules out many non-intrinsic explainable AI methods that rely on apparent or likely reasons, or post-hoc rationalisations of opaque reasoning. Prompt engineering techniques like chain-of-thought reasoning simply imitate statistically likely reasoning and does not provide a trace of the actual computations that led to the decision. Furthermore, without intrinsic explanation, developers cannot eliminate the risk of the use or unintentional inference of protected attributes that should never be considered in a decision. No mainstream GenAI tool rendering its own conclusions independently can provide such faithful explanations that may be required under OIA s23.

Box 5.1.2. Research

Algorithms and AI that generate data that support decision-making at the system level or are involved in intervention development itself. These models do not need to be tied to a specific intervention; this category includes models generating routine data that measures the general impact of policy. This category also excludes the use of AI tools outside of evidence generation. For example, using generative AI for the general collation of research automates part of the routine operations of staff, and is thus administrative in nature.

Here, data from individuals are used in aggregate. The Privacy Act allows the use of already collected personal information for statistical or research purposes if individuals cannot be reasonably identified from the outputs of the research. This exception also enables research techniques like the Integrated Data Infrastructure and synthetic data to be used outside the purpose of collection, if confidentiality can be guaranteed.

Outputs of these systems form part of the evidence base for system-level decisions. Thus, governance around such systems is primarily to assure robust and impartial advice is given to those who make these decisions (e.g. Ministers, Crown agent boards, management). Nevertheless, other ongoing capability maintenance and other risk controls like QA will still be necessary to ensure human-factor risks remain low.

Examples:

- **Research algorithms:** rules-based qualitative coding, CBAx cost-benefit analysis,
- **Supervised models**: StatsNZ net migration nowcasting, building consent data automated coding; Reserve Bank GDP nowcasting proof-of-concept
- **Unsupervised research models**: clustering cohorts of interest to target interventions
- **Goal-driven optimisation research models**: MoT's Monty simulation (co-evolutionary computation), PHF Science's ALMA digital twin (large population model), MoE school bus route optimiser (possibly traditional combinatorial optimisation)
- **Research large language models**: thematic analysis of qualitative data like customer feedback and statutory consultation, orchestrating LLMs that perform analysis and predictions.

Challenges:

- **Transparency of models:** Requests for this information are handled by the more general section 12, where information can be withheld by any exemption. Objective data and evidence generally are not withheld to the same extent as advice and opinion.
- **Trustworthy policymaking**: The subject matter surrounding research AI may be conceptually inaccessible to the public, if not inaccessible due to conclusive or good withholding grounds. Citizens, and accountable decision-makers who also may not have the same degree of expertise, place a unique trust in these specialists to provide "comprehensive, objective and balanced" policy advice (as characterised by DPMC). Given the reduced accessibility, increased breadth of impact, and the presence of context-agnostic technical challenges from Section 4, fostering public trust in research AI should still be sought through robust quality assurance processes and disclosure.

Box 5.1.3. Administration

Algorithms and AI that automate (or otherwise significantly support) core functions of government staff in routine operations that does not otherwise have a direct effect on individual's rights or interests. The execution of administrative systems is still critical to operational integrity and requires performance monitoring to minimise risk. A function is only considered 'core' if its performance is fundamental to the output of a human-equivalent role. For example, a legal advisor may require performance improvement if they are consistently unable to correctly identify, interpret and summarise facts and case law, but will not for incidental issues like spelling mistakes, poor meeting transcripts, or unengaging presentation visuals. Reapplying this analogy brings document summarisation in scope, but spell check, meeting transcription and media generation out. This definition is fundamentally risk-based and self-assessed; it is up to agencies and their risk appetite to determine whether a system is worth monitoring.

Examples:

- **Administrative algorithms**: business rules e.g. customer relationship management, internal enterprise rules e.g. payroll calculation, document classification as per Protective Security Rules
- **Supervised administrative models**: DOC object identification in aerial camera imagery, MPI face redaction and event flagging in fishing vessel footage, StatsNZ automated coding in building consents data
- **Unsupervised administrative models**: fraud and abuse event flagging, where no definitive decision is made
- **Administrative generative AI**: customer service chatbots, assistive web search, customer call transcription, customer call real-time knowledge discovery, document creation e.g. policy documents, business processes.

Challenges: as for the specific type of algorithm used, e.g. varying degrees of automation bias by technique; user acceptance testing is still advisable. Use of these kinds of AI may still be contentious with the public, e.g. staff rigidly following AI recommendations without an element of human discretion, so process controls remain relevant.

5.2. Action: Overlay the two categorisations to refine algorithm and AI guidance

Other jurisdictions, such as the EU and Canada, along with the current Algorithm Charter, adopt a risk-based approach to classifying algorithms and AI. New Zealand's approach has been criticised for being too flexible to be practical, as it depends on agencies to self-assess the material impact of algorithms while offering little guidance through precedents. More effective taxonomies have been developed by organisations such as the OECD. Their Framework for the Classification of AI Systems offers 31 different criteria for classifying AI systems, grouped into 5 dimensions. Although this framework may be suitable for the international, sector-spanning

environment that the OECD operates in, much of the criteria can be answered similarly, some even irrelevant, for various government applications.

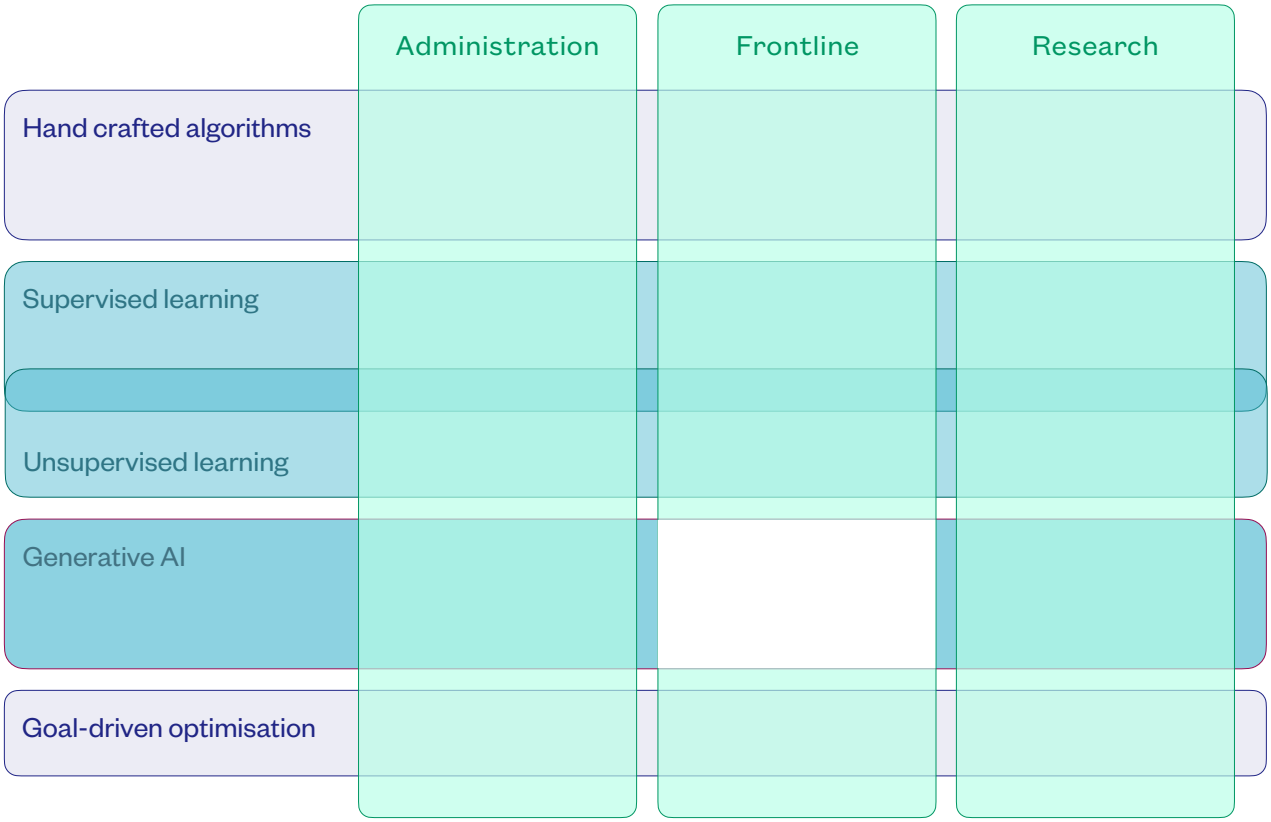


Figure 4: Diagram overlaying the technical categories outlined in Section 4 (symbolic AI, supervised learning, unsupervised learning, goal-driven optimisation and generative AI) with the contextual categories outlined in Section 5 (front-line, research, and administration). Frontline generative AI systems are currently shown as a gap deliberately given the limitations of generative AI systems in fulfilling transparency obligations.

A two-way taxonomy, as shown by Figure 4, offers sufficient practical specificity for monitoring and governance within the New Zealand context. One axis delves into just enough technical detail to group similar monitoring and evaluation approaches into a category. The other axis is based on the environment and risk profile in which the system operates. When new technologies emerge, guidance along the contextual axis should already be in place to help agencies navigate their existing legal obligations, fostering an environment in which they can innovate safely. Appendix 2 demonstrates how this taxonomy can be used to categorise existing and potential AI use cases, providing a visual method to conceptualise how AI is used and could be used in government.

6. The NZAIM unifies technical guidance, while supporting frameworks can be re-engineered

As this paper argues, the existing legislative framework already offers the foundation to enforce best practices where necessary. The focus of the recommendations is on re-engineering supporting frameworks and guidance. Figure 5 illustrates a proposed, more streamlined guidance ecosystem. Normative commitments around the trustworthy and considerate use of any information, especially system-specific commitments in the Algorithm Charter and Ngā

Tikanga Paihere, as well as partnership-specific commitments promoted by Te Kāhui Raraunga, can be refactored around the existing Data Protection and Use Policy (DPUP). A single, enhanced, and mandated DPUP+ will serve both as a blueprint for agencies to develop a high-performing, trustworthy, and respectful data culture, and as a benchmark to independently evaluate an agency's adherence to best practice.

Operational considerations related to the development and end use of AI systems can be unified within a New Zealand Artificial Intelligence Manual and streamlined Responsible AI Guidance sections. Both should be aligned with the principles of the Public Service AI Framework to clearly demonstrate how this strategy is operationalised. An NZAIM will incorporate (rather than simply reference) existing obligations and guidance on accessibility, security, cloud risks, record management, and privacy.

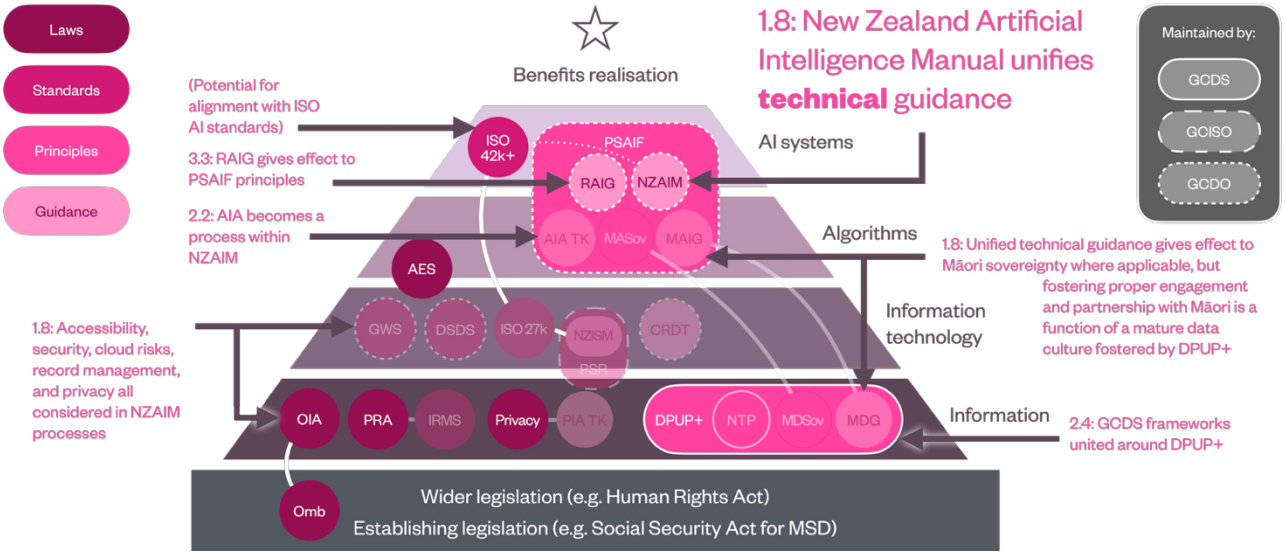


Figure 5: Proposed state for a streamlined guidance ecosystem, refining the roles of the GCDS and GCDO into single frameworks at each abstraction layer: DPUP+ and PSAIF respectively.

This streamlined ecosystem clarifies the overlapping roles of the GCDO and the GCDS by defining – but functionally linking – the ethical "why" and the operational "how". Abstracting the "why" at the most fundamental level provides a lasting foundation for any "how", both now and in the future.

Acknowledgements

Special thanks go to Gareth Derby (Office of the Ombudsman) for providing AI-specific guidance around the OIA and Ombudsman Act, and to Matt Farrington (THW-VUW), Tom Barraclough (Syncopate), Emma MacDonald and Emma Naji (StatsNZ Centre for Data Ethics and Innovation) for providing feedback. The GCDO was unavailable for comment when contacted. The cover image's background is in the public domain and was photographed by Lucas W via Pexels.

Appendix 1: Summary of challenges identified in each technical (Section 4) category

Legend	*Not applicable*	*Dependent*	*Manageable*	*Difficult*

Build: Challenges addressed during the design and training (if data-driven) of the system					
Challenges	**Hand-crafted**	**GDO**	**Supervised**	**Unsupervised**	**Generative**
Output-outcome misalignment (proxy mismatch)					
Biased, inequitable or homogeneous outputs	Designer responsible	Designer responsible			
Opacity makes it difficult to understand or review		Technique dependent	Technique dependent	Technique dependent	
Input data accuracy assumed					Often not known
Confounding variables biases outputs					
Value chain hard to interrogate			If pre-trained	If pre-trained	
Problem definition difficult				No ‘truth’	Highly generalisable
Target(s) accuracy assumed				No target	Often not known
Output accuracy isn’t validated by design					
Process: Challenges outside the system’s build and use relating to how users and the environment interact with the system.					
Users vulnerable to automation bias					More active controls needed
User hijacking	‘Game’-able	‘Game’-able if live	‘Game’-able	‘Game’-able	Often arbitrary
Use: Challenges that may arise during the operation of the system					
Concept or data drift					
Compute cost		If DL	If DL	If DL	
Hallucination					
Unauthorised data reuse					

Appendix 2: Map of NZ government algorithms and AI under the technical-contextual taxonomy

** Pre-training generative AI systems is not the object of generative AI evaluation, which relies on expert and end-user acceptance in the specific context it is applied in, rather the general performance benchmarking undertaken in pre-training. Technically, pre-training (and prior fine-tuning) uses a combination of self-supervised, supervised and reinforcement learning. While fine-tuning may take place when the context has been established, this is only a means to improving the performance upon evaluation, not the end evaluation of the system per se.*