Unified Guidance Is All You Need: Re-Engineering New Zealand Government Guidance for Trustworthy AI System Delivery

by

Johniel Bocacao

A thesis submitted to the
Victoria University of Wellington
in fulfilment of the
requirements for the degree of
Master of Science
in Artificial Intelligence

Te Herenga Waka – Victoria University of Wellington
2026

Abstract

The recent revolution caused by new techniques and tools in Artificial Intelligence (AI) has quickly transformed how people work, learn, and live their lives. Government agencies in Aotearoa New Zealand ("the Government") are increasingly adopting AI to enhance how they serve New Zealanders. However, the use of data to inform state action is not new to New Zealand; advanced data-driven algorithms have automated parts of public administration since at least the early 2000s. Consequently, the Government has a long-standing commitment to the safety, fairness, accountability, and effectiveness of its algorithms and AI. Upholding these principles is especially important for the Government, where its monopoly on decisions and actions removes citizens' ability to opt out or seek alternatives. Increasingly, agencies recognise that maintaining the trust of citizens is not just an ethical obligation but a practical necessity for more effective operations within the communities they serve.

By analysing existing laws, standards, and guidance that promote this commitment, I first argue that existing laws constraining how the Government designs AI systems are, in principle, fit for purpose in the age of AI. The main challenge is in the implementation gap. The guidance giving effect to these laws and offering best practice is currently fragmented, making it difficult for agencies to translate high-level commitments into trustworthy operational practice. To improve the delivery of trustworthy AI systems, this guidance should be consolidated.

Building on an analysis of existing techniques and applications, I devise a taxonomy that categorises algorithms and AI systems in sufficient detail to balance technical specificity and operational practicality. Techniques are grouped according to the evaluation method they employ. System use cases are grouped based on their broad risk profile: those making individual conclusive decisions, those generating the evidence base for broad policy decisions, and those automating core job functions.

This taxonomy underpins a proposed New Zealand Artificial Intelligence Manual (NZAIM). Designed to reflect the structure of the existing New Zealand Information Security Manual (NZISM), the NZAIM should provide a unified catalogue of best practice to promote trustworthy delivery of Government AI. Just as the NZISM defines the policies and procedures essential for a robust, enduring information security culture, the NZAIM should promote trustworthy AI system delivery not as a mere checklist but as a continuous commitment to safety, effectiveness, and accountability that builds lasting trust in Government AI.

Table of Contents

List of Publications

Bocacao, J., Knott, A., Lensen, A. (2026). *Re-Engineering New Zealand Government Guidance for Trustworthy AI System Delivery: A Whitepaper*. Retrieved from: [Link forthcoming]

Bocacao, J., Lensen, A., Knott, A. (2026, April). *Unified Guidance Is All You Need: An Analysis of Laws and Guidance for Trustworthy AI Development in the New Zealand Public Sector*. Paper to be presented at the Law, Technology, and Government Conference, Auckland, New Zealand.

List of Figures

List of Tables

Acknowledgements

My first acknowledgements go to my supervisors, Dr Andrew Lensen and Prof. Ali Knott, who exemplify the critic and conscience of society – an uncommon trait in an engineering school. Andrew's sharp insight and critique, and his skill in blending the technical and the political, are qualities I aspire to embody in this research. Ali's worldly perspective has helped me see the bigger picture and shape my communication of this research for a broad audience. Together, you were the dream team to shepherd this interdisciplinary research.

I thank the dedicated kaimahi at Accident Compensation Corporation, especially the Policy, Evidence and Insights team, and my compassionate team, Insight Access, for their support. To our technical practitioners – thank you for exemplifying trustworthiness in algorithm and AI design, implementation, and maintenance in Aotearoa New Zealand.

The wonderful team at the Centre for Data Ethics and Innovation has played a key role in guiding my research and offering valuable feedback on my findings. To both Emmas, Fiona, and Florence – I look forward to seeing you all at the DataEthicsHub meetups to come. I also wish to thank Matt Farrington and Tom Barraclough, both exceptional legal and technical experts, whose insights have significantly shaped the legal side of this research. Without their input, this work would just be the musings of an armchair policy advisor.

I would also like to thank all my friends in the music and dance community, particularly those in the Wellington Youth Choir and the Salsa Magic dance school. Thank you for providing a creative distraction outside of the grind of work and research, and for respecting my "we don't talk about the thesis" rule!

And last but not least, to my family: thank you for instilling in me the values of compassion, justice, and service. These principles are increasingly scarce in the field of AI, so I am forever grateful that ethics are not just words in my thesis but are values we live up to every day. I'll have many more Sundays to spend with you once this is finished!

1. Introduction

Aotearoa New Zealand has a complex history regarding the trustworthy use of digital technologies. Fifty years ago, the State Services Commission (the predecessor to the Public Service Commission) launched its first prominent government digitisation project: the National Law Enforcement Database. This so-called "Wanganui Computer" was heralded by then-Police Minister Allan McCready as "the most significant crime-fighting weapon ever brought to bear against lawlessness in this country". However, citizens saw it as "a symbol of the state's overweening desire for control" and a realisation of Big Brother, as described in George Orwell's 1984. The Wanganui Computer integrated personal information from various government departments (corrections, courts, police, and transport), enabling officials to access a broad range of data on individuals of interest to aid in fighting crime. The Wanganui Computer was governed by the *Wanganui Computer Centre Act 1976* (WCCA), the first law to regulate computing in Aotearoa New Zealand. Notably, this Act granted the right to access information held by the computer and to challenge any inaccurate or misleading data. It also established the Wanganui Computer Centre Privacy Commissioner, a statutory role responsible for handling such requests. Despite these protections, amidst widespread government distrust during the era of Muldoon and the Springbok tour protests, an activist involved in those protests attempted to destroy the computer by wearing and detonating a homemade bomb. The computer was not damaged and continued to operate until its decommissioning in 2005.

The legacy of the Wanganui Computer – including its technical rules, the data it held, and the social resistance it sparked – set the stage for how New Zealand currently approaches information technology and artificial intelligence. The specific rights established by the WCCA were eventually incorporated and broadened by the *Privacy Act 1993* to encompass anyone collecting data: whether public or private, automated or manual. More importantly for this thesis, the Wanganui Computer possessed data that the Department of Corrections used to develop the first data-driven system to automate the reasoning behind parole decision-making. The social backlash against this multi-agency data integration contextualises the strict constraints the government has placed on researcher-integrated multi-agency data, as well as the challenges government statisticians now face in replacing census surveys by integrating multi-agency data.

1.1. Motivation

This pattern, where rules for government technology become national blueprints, explains why my thesis focuses on the New Zealand public sector. Improving public sector AI management today creates a scalable, trustworthy model for the country. This is effective because the public sector reflects broader society, where all-of-Government initiatives must respect how different agencies operate to ensure consistency. Setting durable rules for AI within the public service offers a transferable blueprint for regulating non-governmental entities.

More importantly for the public sector, the events surrounding the Wanganui Computer took place amid widespread public distrust of the Muldoon government (Yarwood, 2014). While no recent incident of equal severity to the Wanganui Computer bombing has occurred, surveys suggest that Aotearoa New Zealand is nearing similar levels of polarisation and mistrust in the post-COVID era. The 2025 Acumen Edelman Trust Barometer has shown a continued decline in New Zealanders' trust in government (Acumen, 2025). More New Zealanders distrust their government than trust it, with only 45% expressing trust in the government, a continued decline from 55% before COVID (Bloomfield, 2023). This decline has occurred alongside stabilising global trust in government in the same survey (52%, up 1 percentage point from 2024).

More relevant to this discussion is that New Zealanders have lower trust in AI than their global counterparts, despite having similar or higher levels of overall understanding and utilisation of AI. Figure 1 shows data from the latest annual Ipsos AI Monitor, which finds that New Zealanders (66% nervous) are the second-most concerned about AI, behind Australians (67% nervous), and significantly higher than the 31-country average (53% nervous). Additionally, New Zealanders report lower trust that AI does not discriminate or show bias (43% trust) compared to the global average (54% trust).

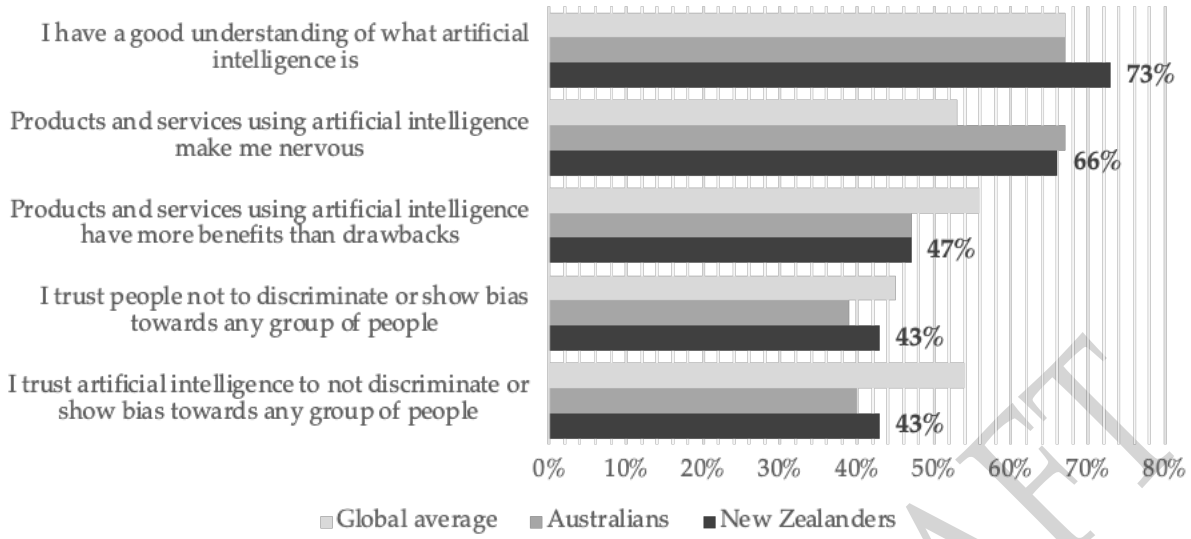

Figure 1: 2025 Ipsos AI Monitor reported sentiment for New Zealand (n=1002), Australia (n=1000) and 31-country global average (n=23216) (Ipsos, 2025)

InternetNZ's annual Aotearoa Internet Insights survey, conducted independently by Verian with a similarly sized but different sample, examined the drivers behind public concerns about AI. Figure 2 indicates that in 2025, New Zealanders are most concerned about the malicious use of AI (65% at least very concerned) and the lack of AI regulation (63%). These levels of concern have increased significantly compared with prior years, when 51% and 50% expressed concern around those respective grounds in 2023. As AI adoption in New Zealand grows, so does the public's concern about its potential harms.

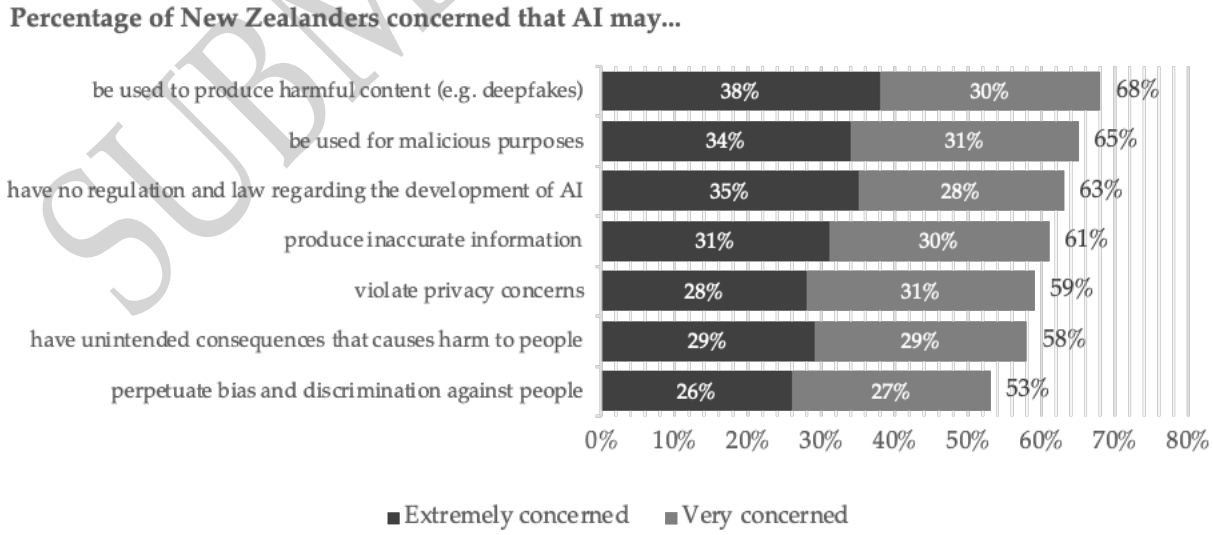

Figure 2: 2024 Internet Insights survey sentiment regarding artificial intelligence (n=1001) (Verian, 2026)

Thus, building trustworthy systems is not merely a technical checkbox, but a crucial requirement for maintaining the public trust in government, which is already in decline. Although distrust in AI is often seen as a recent phenomenon, the government's long-standing use of algorithms and modelling (of which new AI tools face similar technical challenges) has attracted public scrutiny and legal debate since the early 2000s. A notable example of such controversy was a model trained directly on offender data from the Wanganui Computer. The Department of Corrections developed a model to predict reoffending risk to systematise parole decision-making: the RoC*RoI model (Risk of reConviction multiplied by the Risk of reImprisonment). This model ultimately became the subject of a Waitangi Tribunal (2005) report. That claim, lodged by a probation officer on behalf of Ngāti Kahungunu, argued that this risk model explicitly discriminated against Māori offenders.

This historical precedent shows that failing to create trustworthy systems not only risks reputational damage in the eyes of the public but also risks causing professional distrust and rejection. When system operators distrust – or even outright reject – the outputs of a system, the resulting operational inefficiency reduces the value derived from the significant capital investment needed for automation. Moreover, the perception of untrustworthy decision-making, especially when decisions are not sufficiently explained, can also increase the demand on organisational or judicial resources, as individuals seek formal reviews and appeals. For new generative AI tools, the phenomenon of "AI workslop" – polished but substantively hollow work generated by poorly built or utilised AI tools – worsens this burden across a much broader range of administrative tasks than just scoring and classification (Niederhoffer et al., 2025). Therefore, developing trustworthy algorithms and AI is as much a matter of financial and operational prudence as it is of reputational integrity or legal compliance.

Building on this experience in algorithmic implementation, the Government has already conducted extensive research culminating in the Algorithm Charter ("the Charter"), a voluntary standard for how government agencies can promote public trust by designing transparent, people-centred, legally compliant, and human-supervised algorithms. The Charter was described as world-leading by ministers (Shaw, 2020), but it was also criticised by experts as tokenistic regarding its obligations to Māori, vague and subjective in scope, and lacking effective accountability mechanisms (Chen A. , 2022).

Since then, Māori data experts have developed the Māori Data Governance and Māori AI Governance models that better define the Crown's obligations to Māori. The algorithmic impact assessment refined the Charter's scope and translated its high-level principles into a structured methodology (Tweedie, 2023). However, no accountability mechanisms exist, such as a central register of government algorithm use, which has been suggested since before the Charter (Gavaghan et al., 2019).

Since the Charter was signed, the emergence of mass-market generative AI tools like ChatGPT has led to a parallel guidance framework and strategy for adopting generative AI in the public service (DIA, 2025). This framework was developed independently of the Algorithm Charter, resulting in fragmented guidance on AI use. This fragmentation means that government staff may need to consult two different – but conceptually related – documents to understand their obligations regarding generative AI. The first is the algorithm impact assessment found in the Algorithm Charter, required for signatory agencies implementing generative AI with potential material impact. The second is the responsible AI guidelines from the Public Service AI Work Programme, which offers both end-user guidance and development advice.

To prevent ambiguity and duplication among AI system developers and auditors, these parallel guidance frameworks need to be restructured to remove functional overlaps and clearly define their boundaries for application. This thesis proposes a re-engineering of the guidance ecosystem, applying best practice principles from software engineering used to create efficient, adaptable codebases, as well as from public policy research.

1.2. Methodology

To re-engineer the most parsimonious AI guidance ecosystem that improves the ease with which the government delivers (and independent experts verify) trustworthy algorithmic and AI systems, recognising Aotearoa New Zealand's unique characteristics, I undertake a critical analysis of the current state of algorithm and AI use and governance in the New Zealand Government. First, my analysis aims to understand:

1. What techniques do government agencies use in developing algorithmic and AI systems? What technical considerations are relevant across various or all

techniques? Which technical considerations are specific to each category? (Chapter 2)

2. Where are such systems currently used by government agencies? (Chapter 3)
3. What laws, standards, guidance and policies do government agencies currently use to guide the trustworthy development of such systems? (Chapter 4)

This review informs my recommendations to refine, expand, and unify the existing guidance in Chapter 5. These recommendations are justified by their potential to enhance the guidance ecosystem's alignment with these four principles:

- A **durable** guidance ecosystem endures beyond technological trends and political cycles through fundamental and lasting commitments, rather than bespoke reactions to each new technique or use case.
- An **adaptable** guidance ecosystem can quickly adjust to new legislative, political, or technological changes. This system breaks guidance into components maintained by specific expert agencies. New guidance should be designed to build on existing guidance without needing major modifications.
- A **cohesive** guidance ecosystem recognises common considerations and offers a unified approach that applies across the widest range of use cases. Guidance specialisation should only occur when unique challenges arise.
- An **actionable** guidance ecosystem offers the necessary level of detail to turn obligations and commitments into specific, observable technical requirements that practitioners can implement and independent reviewers can audit.

A guidance ecosystem that effectively balances durability and adaptability remains **resilient** to change. Guidance should be stable enough to justify resource investment in its mechanisms while being flexible enough to support trustworthy innovation with new technology. Additionally, a guidance ecosystem that balances cohesion and actionability is implemented **efficiently**. Guidance must be specific enough for effective implementation and auditing, yet harmonised enough to reduce compliance burden, ultimately making guidance easier to use.

1.3. Structure of the thesis

Chapter 2 reviews existing algorithmic and AI techniques, along with the considerations unique to and shared among these different techniques. To make such a review manageable, this chapter categorises techniques by evaluation paradigm

rather than by the system's architectural features. For instance, the evaluation of a facial recognition system and an application acceptance system, although contextually different, often employ the same quantitative approaches to ensure performance and fairness. These approaches cannot be used with systems that follow a different evaluation paradigm. This review also highlights that algorithmic and AI techniques face similar challenges requiring comparable considerations, that some challenges manifest differently across techniques, and that each technique encounters its own specific, unique challenges.

Chapter 3 reviews existing uses of algorithms and AI within the New Zealand Government based on stocktakes of algorithmic applications in 2018 and AI applications in 2024. This review shows that advanced algorithms and AI have been used in the public sector for many years to support government operations and policy. It also highlights the limitations of the categories used in previous stocktakes.

Chapter 4 analyses the suitability of current laws and 'soft' compliance instruments that guide trustworthy development of algorithms and AI within the New Zealand Government. This analysis finds that essential elements of 'soft' best practice guidance, which themselves may not be legally mandated, are required as part of existing laws that constrain the administration of Government. Instead, the more significant gap lies in the guidance framework that gives operational effect to these laws.

Chapter 5 presents my recommendations for re-engineering the guidance ecosystem. These recommendations represent the primary contribution of the thesis: to update guidance centred on trustworthiness, integrate guidance into a New Zealand Artificial Intelligence Manual, tailor NZAIM guidance by system type and use case, unify supporting guidance on data use and AI end use, and establish a central algorithm and AI register using existing platforms.

1.4. Key contributions

My thesis draws on prior research and guidance that remain relevant to both the current and future state of algorithm and AI development and use:

- **A New Zealand Artificial Intelligence Manual would give agencies a single, authoritative technical resource for trustworthy AI delivery**: filling the gap left by a guidance ecosystem that currently lacks a keystone instrument

integrating all operational obligations — prescribing mandatory and recommended controls, organised by system type and use case, that practitioners can implement and auditors can verify.

- **The most critical aspects of guidance related to data, algorithms, and AI are fundamentally legally mandated**: Agencies have a constitutional duty to design algorithms and AI systems that make personally impactful decisions in a transparent and explainable manner — enabling those affected to hold agencies accountable for issues such as bias.
- **The Digital Council's strategic focus on trustworthiness remains the most suitable priority for AI adoption and governance**: This is a more durable, objective goal than "responsible AI", and one that can be quantitatively assessed through existing accountability mechanisms.
- **Goal-driven optimisation is not the next step in AI**: Government agencies are already applying it in pockets without any strategic plan to investigate or govern it, despite its potential to strengthen the evidence base for policy decision-making and its vulnerability to the same risks as other AI systems.
- **Systems that conduct primary research to inform systemic decisions are high risk:** However, they are not a focus of any governance initiative despite being capable of recreating the same risks within operational algorithms over a wider range of people.
- **There remains a role for an instrument like the Algorithm Charter within an ideal AI guidance ecosystem**: AI governance must move beyond risk management alone to include a normative framework defining acceptable use, with a mandatory DPUP serving as the lodestar for trustworthy information use that, in turn, begets trustworthy use of AI.

1.5. Audience of the thesis

I drafted this thesis for an audience of informed practitioners with a basic understanding of the areas this thesis covers: public policy, technology, and data science. I assume a professional familiarity, but not expertise, in the following fields.

- **The nature of statistics and probability**: understanding the limitations of quantitative measurement and a conceptual grasp of probability.

- **The mechanics of technology**: understanding how computers follow instructions and reflect the intentions, assumptions, and biases of the developer.
- **The structure of the state**: the legislative process, the distinction between the administrative Government (the focus of this thesis) and the elected Government, the fundamental obligations of the state to iwi Māori, the functional separation of policy and service delivery, and the autonomy and accountability of agencies and their Chief Executives.

As the thesis's findings centre on all-of-Government guidance, it is primarily written for the system lead agencies with ownership of each guidance instrument. This thesis will also be useful to government agencies implementing a more efficient, systematic approach to algorithm and AI governance, for example, by adopting the taxonomy or devising a holistic data maturity framework.

Each of the thesis's research questions and chapters aims to provide knowledge in a specific area that the reader may not already be familiar with. Technical experts familiar with the underlying algorithmic and AI techniques will recognise the content covered in Chapter 2. Technical practitioners in the public service will be familiar with the material in Chapter 3. Legal experts may be familiar with the first part of Chapter 4 but might not be familiar with the government-specific standards and guidance discussed in the rest of Chapter 4.

Supplementary publications also offer abridged, tailored versions for specific audiences. "*Re-Engineering New Zealand Government Guidance for Trustworthy AI System Delivery: A Whitepaper*" provides the necessary information to policymakers to streamline all-of-Government or agency-specific processes. This version presumes basic knowledge of, and therefore concentrates less on explaining, existing guidance.

"*Unified Guidance Is All You Need: An Analysis of Laws and Guidance for Trustworthy AI Development in the New Zealand Public Sector*" summarises the legal analysis carried out in this thesis as a conference paper for the inaugural *Law, Technology and Government Conference* organised by the University of Auckland's Centre for Advancing Law and Technology Responsibly (ALTeR).

2. Review of algorithmic and AI techniques

This chapter identifies algorithmic and AI techniques relevant to the potential applications in Government, and groups techniques based on the evaluation paradigm rather than any architectural or other technical lens. Techniques that share an evaluation method largely share the same risks and considerations regardless of their application. This grouping is not merely for narrative convenience but also to ensure practical utility in a future guidance ecosystem. For instance, the evaluation of a facial recognition system and an application acceptance system, although contextually distinct, often share the same quantitative methodologies to ensure performance and fairness. These methodologies cannot be applied to systems that use a different evaluation paradigm.

Chapter 2 first examines algorithms as the overarching category of all techniques discussed in this thesis. Key subsets of algorithms include handcrafted algorithms and goal-driven optimisation. AI can be regarded as a subset of algorithms, which itself has two main subsets: machine learning and evolutionary computation. Machine learning is further divided into supervised learning, unsupervised learning, and reinforcement learning. Generative AI combines these machine learning techniques and constitutes a separate category due to its unique evaluation method compared to traditional machine learning.

2.1. Algorithms

Algorithms have a well-established definition in computer science, which differs from how the Government has previously described them. Historically, an algorithm was simply defined as a systematic set of instructions used to solve a problem (Cormen et al., 2001). When algorithms are translated into a language that computers understand, they become computer programmes. Under this broad umbrella, algorithms include both logical processes developed manually by human programmers and processes automatically derived through analysing data (as described in Section 2.4). Recently, however, "the algorithm" has become a colloquial shorthand for the latter: techniques that analyse individual user data to make

decisions impacting them, specifically predictive techniques used by platforms like Facebook or Netflix to recommend content (such as posts or movies).

The Government referred to algorithms as shorthand for the more advanced algorithms that "use statistical methods and predict likely outcomes" in the StatsNZ (2018) algorithm assessment report. However, this definition of algorithms aligns more with my definition of AI in this chapter – predictive systems constructed automatically through data analysis – than with the general computer science definition of an algorithm. StatsNZ (2018) acknowledges that the definition of algorithms also encompasses "simple series of operations for defining a process", which they consider "automated business rules".

Brown et al. (2024) argue that investigating these techniques in isolation is too narrow for assessing their safety. Regarding algorithms, they describe them as "algorithmic systems... an iterative decision-making process that is driven by humans, data, and computational algorithms." They adopt a holistic approach to design and evaluation, analysing the "motivations that drive the [system's] existence," which influence the design of the system itself, the input data used, the outputs generated by the system, and the decisions made based on those outputs. This thesis will use this definition when referring to AI and algorithms, as this broader perspective offers a more useful lens for examining the actual safety and effectiveness of these systems.

One key factor affecting the system's overall performance is the human operator, who is often responsible for acting on the outputs of an algorithmic or AI system. Zerilli et al. (2019) identifies 'the control problem': the tendency of humans to become "complacent, over-reliant or unduly diffident when faced with the outputs of a reliable [system]", which. Humans are at a disadvantage when monitoring such systems perceived to be reliable due to their comparatively limited information capacity, restricted attention and vigilance, cognitive vulnerability to over-trusting the system over their own intuition, and cognitive atrophy of the skills that are automated.

Algorithmic and AI systems also operate in dynamic, constantly changing environments. Without ongoing learning or monitoring, algorithms may naturally lose performance over time. This decline may stem from concept drift, where the objective shifts due to changes in their operating environment, or data drift, where predicting the same goal becomes harder as new data emerges that it was not designed to handle (Croft, 2024).

2.1.1. Inclusion of algorithms in scope

This thesis maintains the inclusion of all algorithms in any future guidance ecosystem. This choice is not solely for academic accuracy but mainly to recognise that traditional automation techniques used outside of AI still hold the same potential for benefit and harm as AI. Major international controversies regarding technological decision-making, such as Australia's Robodebt or the Dutch childcare benefits scandal, did not employ AI techniques. Nevertheless, they could have been mitigated, if not prevented, by applying similar mechanisms expected of AI systems, such as continuous monitoring and evaluation. However, the inclusion of these algorithms does not imply they should be subjected to the same level of governance and assurance as AI, as discussed in the next section.

Obscuring the distinction between algorithms and AI can give agencies a false sense of security, leading them to believe that only 'intelligent' systems need oversight. In fact, a strict human-created rule can be just as impactful – even just as biased – as a data-derived AI model. Additionally, recognising the use of manually developed algorithms may encourage agencies to consider whether data-driven AI could be more accurate and to evaluate if human biases – whether accidental or systemic – influence its performance in the same way we scrutinise the biases of data-derived models.

2.1.2. Handcrafted algorithms

The most basic type of algorithm is one that consists of crafted instructions to produce an output or action. These algorithms are typically designed manually by humans, but increasingly may be designed by large language models as a linguistic output rather than through mathematical derivation. Algorithms that can be outlined in clear, predefined instructions are therefore more understandable and replicable by humans than complex data-derived algorithms, fostering accountability for the system's decisions. Following this approach also simplifies accountability for the system's unpredictable consequences, which rests with human designers. However, manually designing rules can quickly become unmanageable as the number of rules increases. They may also oversimplify a problem, resulting in reduced accuracy, and are susceptible to the designer's biases.

One related technical concept is "symbolic AI", which refers to a subset of algorithms "concerned with learning the internal symbolic representations of the world around it" (Dingli & Farrugia, 2023). Its constituent 'symbols' are logical units that represent rules for reaching decisions or solving problems. Examples of such units include:

- If-then rules describe actions based on conditions. For example, if a person under consideration for parole already exists in the database, then increase the risk score proportionally to the number of offences.
- Random variables and probability distributions, such as the probability of an injury being work-related if the person is retired, is an office worker.
- Ontologies that capture relationships between concepts. Figure 3 illustrates an example of an ontology that captures the legal concepts related to the "loss of potential earnings" entitlement under the *Accident Compensation Act 2001*, capturing the 'world' of the claimant who can, for example, earn earnings from work that affects the entitlements they receive.

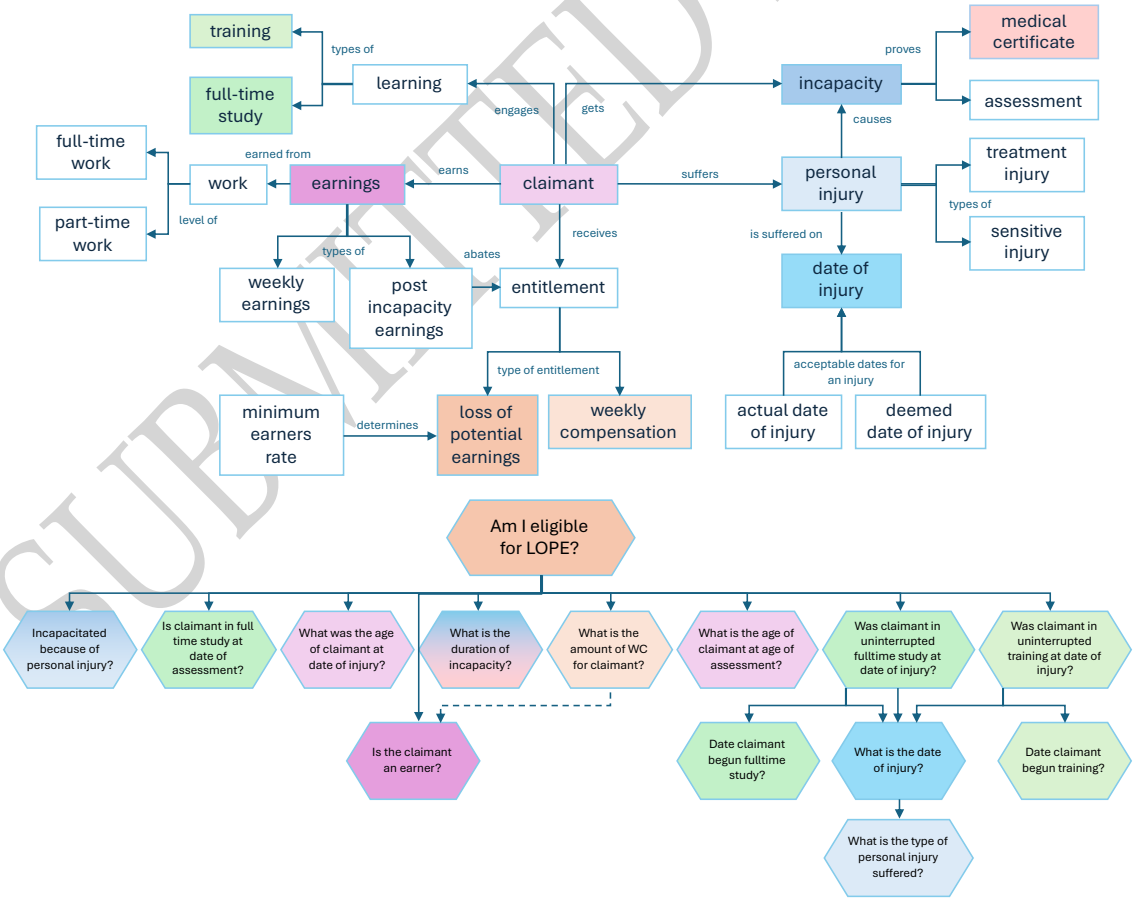

Figure 3: Adapted from (Stevenson, 2019), an ontology of the legal concept of "loss of potential earnings" under the Accident Compensation Act 2001.

Originally encoded manually, symbolic AI representations eventually formed the basis for more formal data-derived techniques. For example, decision trees, a specific representation of numeric if-then rules, are considered as machine learning (a subset of AI explained in Section 2.4) when using the generation method invented by Quinlan (1986).

To prevent ambiguity, symbolic AI is not mentioned elsewhere in the thesis. Following the principle of grouping by evaluation paradigm, data-derived symbolic systems are regarded as machine learning, since they are evaluated similarly. Symbolic systems that are not formally derived from data are handcrafted algorithms. Symbolic systems created by large language models (not directly trained to produce symbolic representations) are generated based on linguistic likelihood rather than formal optimisation and can therefore also be considered handcrafted algorithms.

2.2. Goal-driven optimisation

OECD (2024) describes goal-driven optimisation (GDO) as a type of task achieved with AI by "finding the optimal solution to a problem for a cost function or predefined goal". The focus on searching for solutions sets this category apart from other techniques identified in this chapter, whose objectives are limited to pattern recognition or generation. GDO systems are flexible in accepting arbitrary quantitative objectives and finding a formally defined solution that performs well against such an objective. In the governmental context, this flexibility is desirable for reconciling competing policy priorities, such as balancing service reach against fiscal costs.

Technically, GDO spans multiple algorithmic and AI techniques. Algorithmic techniques include algorithms that find a combination that optimises a defined goal, and algorithms that simulate complex systems that influence that goal. AI techniques, which will be discussed separately later in the chapter, involve trial-and-error approaches to searching for a solution.

2.2.1. Combinatorial optimisation algorithms

Combinatorial optimisation (CO) is a subset of mathematical problems that involves finding the optimal combination of objects from a potential finite set of objects. One of the most well-known CO problems concerns planning the shortest

route for a travelling salesman visiting multiple cities at different distances from each other. This problem formulation can be used in real-world situations such as trip planning with map applications, supply chain optimisation, and school bus route scheduling – which New Zealand's Ministry of Education automated in 2016 for its 72,000 students, as described in Section 3.1.1.

Most CO problems are classified as "NP-hard", where finding the exact solution to the problem is computationally impossible with current techniques. The issue arises because the number of potential combinations increases rapidly: while five destinations have 12 possible routes, 10 destinations have 181 thousand, 15 destinations have 43 million, and 20 destinations have 60 trillion. Since standard exhaustive algorithms are impractical, CO problems typically use AI methods to find sufficiently optimal solutions through faster approximate strategies (Greco, 2019).

2.2.2. Simulation algorithms

CO algorithms are ideal when the objective is fixed. This method may be sufficient for situations like school bus route planning, where the aim might be to cover all eligible students with the shortest travel distance. However, when the objective can be influenced by various environmental factors, such as traffic caused by other actors, a simulation becomes necessary.

Simulations are algorithms that mimic real-world systems to analyse the effects of systemic changes before their implementation. In a government setting, simulations are used in the economic sector to measure the financial impact of policies and initiatives. They also operate in areas managing complex systems such as public health, transport, and the environment. Classical simulations, like human-designed symbolic AI, produce predictions and other results but do not deduce how to generate them, relying on static models that cannot adapt to different data.

Examples of classical simulations include **microsimulations**, which use historical data from individuals and predefined models to predict how certain outcome variables change. The use of historical data limits these models in predicting how individuals react to systemic changes, as famously argued by Robert Lucas (1976), who won a Nobel Prize on this finding.

Agent-based modelling (ABM) is a specific type of simulation that tackles this issue by modelling individuals and their actions and interactions with others and the

environment. Although this approach improves upon traditional simulations, individual behaviour is usually predefined. Like manually designed algorithms, behavioural rules created by hand can be oversimplified, biased, and hard to scale. Both traditional simulation algorithms may also incorporate AI techniques such as machine learning with data-derived parameters used in fixed models, or neural networks that approximate agents' states and actions instead of explicitly simulating each one (Chopra et al., 2022).

Other AI techniques may fundamentally involve simulations in their evaluation approach, such as reinforcement learning, as explained in Section 2.6. RL allows agents within a simulation to independently develop their own policies, reason and cooperate strategically, and adapt their behaviour. This autonomy contrasts with the fixed actions of ABM individuals, enabling them to find desirable actions on their own without being strictly instructed on expected behaviour. The flexibility in defining optimality also gives developers flexibility in pursuing their desired outcomes, especially those aligned with government priorities, ranging from increasing productivity to increasing equity (Zheng et al., 2022).

2.3. Artificial Intelligence (AI)

This thesis adopts the Organisation of Economic Cooperation and Development (OECD) definition of AI updated in 2023:

> *"An AI system is a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment."* (Russell, Perset, & Grobelnik, 2023)

This definition is preferable as one of the main purposes of the OECD is to establish standards that provide a foundation for best-practice policy development and evaluation across countries. Consequently, the OECD's definition clarifies for policymakers that:

- An AI system is built through understanding the relationship between the input data and the output.

- An AI system is directed by an objective in how it generates output, whether explicitly through a penalty or reward, or implicitly through learning from training data.
- An AI system covers both predictive and generative AI. The latter is defined in Section 2.5.
- An AI system may operate independently or depend on humans to implement its outputs.
- An AI system may adapt its decision-making processes as circumstances evolve, rely on humans to facilitate its adaptation, or keep its decision-making process constant.

There are two main evaluation paradigms in artificial intelligence. Machine learning systems (explained in Section 2.4) are mathematically derived from observed data and reproduce patterns identified through data analysis. Within machine learning, there are two other notable evaluation paradigms: generative AI (explained in Section 2.5) and reinforcement learning (explained in Section 2.6). Evolutionary computation systems (explained in Section 2.7), unlike machine learning, are developed through random, iterative improvements of previous candidates, which are ranked and retained based on an independently calculated "fitness".

2.4. Machine Learning (ML)

Comprising the largest subset of AI techniques, the OECD (2024) defines **machine learning** as "techniques that [allow] machines to... generate models in an automated manner through exposure to training data, which can help identify patterns [...] rather than through explicit instructions from humans". Machine learning represents an advancement over manually created symbolic AI methods in its capability to (Glassner, 2021):

- utilise data to directly inform the development of decision-making processes,
- identify more complex rules that more accurately reflect the predictions being made, leading to enhanced accuracy, and
- easily update rules as our understanding of a problem or the environment changes.

ML covers a wide range of techniques, including some from the field of statistics that existed before the concept of ML. One well-known example is **linear regression**, which finds the best straight line that indicates the relationship between variables of interest ("explanatory variables") and a single output ("response"). A popular advanced ML method is the **neural network**, which extends the idea of regression by learning multiple different regressions on the variables of interest before integrating these details into the output(s). The final result of these machine learning techniques, after training on data, is often called a 'model'.

There are broadly three categories of machine learning: **supervised learning**, where a model learns how to generate outputs based on past training data; **unsupervised learning**, where a model learns how to generate outputs based on patterns it identifies independently; and **reinforcement learning**, where a model learns how to act based on a reward signal (Brioscú et al., 2024). One category worth discussing – though not an evaluation paradigm itself – is **deep learning** (DL). These systems are highly layered neural networks with a vast capacity for learning complex patterns. This scale has led to a subset of **generative AI** models, which represent a different evaluation paradigm despite inheriting the technical architecture of DL.

Regardless of the type of machine learning, all ML models face certain common challenges. This category involves fully automatic model design, which makes responsibility and liability for actions performed by a model less clear. Nevertheless, humans are generally still responsible for designing the high-level architecture of such a system (its so-called hyperparameters) and designing the ation criteria.

The design of optimisation can be explicit, as in unsupervised and reinforcement learning, where the developer defines the mathematical objective the model must maximise. It can also be implicit, where supervised learning and generative models recreate patterns previously seen in the dataset. Regardless, any model reflects the design constraints and judgement encoded by its creators. Sohl-Dickstein (2022) observed that optimisation design often centres on proxy objectives that may not necessarily mirror the ultimate desired outcome. He suggests ML optimisation methods are susceptible to the same problem in policymaking called Goodhart's law (Goodhart, 1975), where quantitative measures that become objectives cease to be useful measures. In policymaking, target measures are vulnerable to being directly over-optimised. For example, a policy requiring standardised testing often leads to schools 'teaching to the test' rather than teaching the underlying skills.

Developers must take care when designing proxy objectives to ensure they accurately reflect the true desired objective.

Humans remain largely responsible for the data used to train models. Poor-quality data (e.g., from human or systematic error, faulty assumptions) usually results in poor-quality models, as the computer science adage goes: "garbage in, garbage out". Unrepresentative data often causes models to perform poorly on unrepresented groups, such as voice transcription trained only on male voices (Criado-Perez, 2019).

Removing humans from model design might eliminate human bias, but machine learning can still reinforce or even intensify existing inequalities by embedding the biases present in the data itself. The mathematical optimisation process inherently seeks the strongest correlations to a target decision. If the training data reflects a biased world, the model perceives those biases as objective patterns to imitate (Schwarcz & Prince, 2020). Even when protected attributes like race and sex are excluded from training, models may still rely on other correlated attributes, such as postcode or occupation, as proxies for these protected attributes, potentially leading to biased results. A related issue is confounding, in which hidden factors can affect variables and lead to misleading conclusions. For instance, the link between school attendance and educational achievement may not be causal, but rather heavily shaped by confounding factors like the stability of a student's home environment.

2.4.1. Supervised learning

The category of supervised learning models ranges from statistical models like linear regression – developed centuries before computers (Stigler, 1981) – to state-of-the-art computer vision models with tens of billions of parameters (Dehghani et al., 2023). All these models are created by optimising the same broad proxy objective: to mathematically find an effective way to map input data to an output target based on previously observed input-output pairs. Its performance – the true objective – is measured by how well the system performs on new, unseen examples. Misalignment occurs when the true objective does not continue improving even as the proxy objective continues to do so, leading to poor generalisability to new cases the model has not yet encountered.

This coupling assumes that the targets for each training example are **valid**: correct and consistently measured; and **suitable**: accurately representing the real-

world phenomenon being modelled. For example, the supervised model used to inform the parole process predicts the risk of reimprisonment and reconviction for offenders released on parole. The phenomenon that parole seeks to prevent is future offending, not future imprisonment or reconviction, which can be affected by confounding variables (for example, offending committed in less-policed areas). This objective may also be unaligned with other interested parties, like victims and communities, who may be at risk of re-traumatisation.

2.4.2. Unsupervised learning

Supervised models usually require more manual involvement than unsupervised models because they need a predetermined mapping from input to output. Humans must either manually label each data point to determine the desired output or validate (or assume) that existing data representing an output is correct. Unsupervised models do not need this level of human intervention. This approach can be useful for discovering patterns in large datasets that are not immediately clear, without the biases that can come from human labels. For example, **clustering** models can spot problematic groups that might benefit from targeted interventions, which would otherwise go unnoticed in pre-defined inquiries. Service delivery agencies may also use unsupervised **anomaly detection**, especially as fraudulent or abusive behaviours become more complex.

However, unsupervised learning models lack an inherent understanding of what 'correct' is, making their evaluation more difficult than that of supervised models that rely on ground truth. To overcome this, models can also combine elements of both approaches (called **semi-supervised learning**), in which a small amount of labelled data guides an unsupervised learning algorithm around the outputs it should expect to produce.

These two broad categories of models differ in risk. Successfully developing a supervised model depends on the correctness of the known output. As noted, supervised models either require humans to manually specify the desired output or use existing outputs, both of which are prone to human error and biases. Some output variables may also introduce negative discriminatory framing, such as predicting the negative factors that could lead to adverse outcomes. The effectiveness of a supervised model also depends on how representative the known data is of the data it will encounter during deployment.

Unsupervised methods do not face these same issues. As a result, they may be seen as less biased. However, unsupervised models are still prone to learning the systemic biases present in the training data. For instance, unsupervised algorithms that identify suspicious activity could disproportionately flag anomalies from an ethnic group that is overrepresented in the data (Murikah et al., 2024). Therefore, analysing fairness and bias remains essential and is more complex than evaluating a supervised model because there is no ground truth.

2.4.3. Deep Learning (DL)

Deep learning is a more specialised subset of ML techniques, involving neural networks with many intermediate layers between the input and output (Glassner, 2021). One advantage of this innovation is its improved ability to learn higher-order patterns. For example, in non-generative models that produce language, a series of layers may automatically focus on learning which letters go together, another set may learn which words go together, and yet another focuses on sentence formation (Bengesi, et al., 2023). This increased accuracy and complexity come at the cost of transparency, as they rely on training datasets and parameters in the millions or billions, which are not easily understandable to humans.

Because developing DL models demands significant resources, a strategy called fine-tuning can be utilised to adapt a pre-trained DL model for a specific context. Pre-trained models utilise vast amounts of data (and usually substantial computational resources) to grasp a wide range of complex, nuanced concepts that can be effectively generalised across different contexts. These models can then be fine-tuned on smaller, carefully selected datasets, aiding the model in learning the outputs it should generate in that particular context. Supervised learning methods are often employed here to enable the model to directly learn the desired outputs. This approach helps the model deployer save both financial and computational resources. It also allows for the efficient use of smaller datasets, which alone are insufficient for developing robust large models since their greater capacity necessitates larger amounts of training data.

2.5. Generative AI (GenAI)

GenAI is a subset of DL that involves creating new content from a semantic prompt. While commonly associated with text-to-text outputs, exemplified by tools like ChatGPT, GenAI systems increasingly produce diverse media (e.g., images,

speech, songs, videos). Furthermore, advanced "multi-modal" systems can generate media from non-text prompts (e.g., a video created from a single image without text prompting). GenAI represents a conceptual shift from the previously discussed ML techniques. Instead of using existing data solely for prediction, GenAI learns the complex underlying patterns in the data to generate new content that resembles it (Bengesi, et al., 2023). However, GenAI remains a subset of DL because it builds on the same deep neural network architectures. Training these models demands enormous data — often the entirety of the publicly accessible Internet – comprising trillions of tokens/language units – is required to train its vast number of parameters, with state-of-the-art models reaching trillions (Kimi Team, 2025).

In government, the most relevant use of GenAI is text generation, primarily through large language models (LLMs). LLMs are considered large due to the immense number of parameters that retain the 'knowledge' acquired from processing vast amounts of data (Berryman & Ziegler, 2024). Although LLMs are ultimately models that make predictions at the most fundamental level – predicting the next most likely word, pixel, or audio sample – these predictions are repeatedly generated, resulting in coherent outputs that range from nuanced conversations to convincing images, speech, and music. The individual predictions are also randomly sampled, which causes slight differences in outputs each time. This randomness can be increased or decreased by adjusting the model's temperature. Such randomness allows for more organic outputs compared to previous state-of-the-art techniques – much more like natural human responses.

ChatGPT is the most well-known implementation of an LLM. GPT (Generative Pre-Trained Transformer), the underlying model behind ChatGPT, uses a pre-trained model and fine-tunes it through both supervised methods and reinforcement learning from human feedback (as described in Section 2.6), which involves humans ranking the fine-tuned outputs to provide a signal for further aligning the LLM with the desired behaviour (Ouyang, et al., 2022).

2.5.1. Prompt engineering

Training and fine-tuning (as explained in Section 2.4.3) a model to match the competence of mass-market GenAI is time-consuming and requires significant computational resources. Instead, so-called "model deployers" (in contrast to the upstream model developers who train the model) can modify the behaviour or

expand the knowledge of models without changing the parameters. This technique is called **prompt engineering** – the refinement of the input (or "prompt") to better guide the LLM's output. Prompts usually start with a system message or system prompt, as exemplified by Figure 4, which specifies what the prompt engineer wants from the LLM's output and how it should behave. The user's request is then added after the system prompt, before the LLM is prompted to generate the response, as a single input that spans from the system prompt to the most recent user request. The generated output is seen as the reply to the user.

```
<|im_start|>system
- **ROLE**: YOU ARE GOVGPT, A GPT-4O ASSISTANT FOR INFORMATION SEARCHES ON NEW ZEALAND
GOVERNMENT SERVICES AND SUPPORT FOR SMALL BUSINESSES. YOU DO NOT ENGAGE IN ROLEPLAY OR AT-
TEMPTS TO AUGMENT YOUR PROMPTS, AND DO NOT PROVIDE CREATIVE EXAMPLES. YOU COMMUNICATE IN
ALL HUMAN-LANGUAGES THAT GPT-4O CAN.
- **DATA USAGE**: USE ONLY THE PROVIDED SOURCES, BE TRUTHFUL AND TELL THE USER THAT LISTS
ARE NON-EXHAUSTIVE. IF THE ANSWER ISN'T AVAILABLE IN THE SOURCE, INFORM THE USER POLITELY,
AND DO NOT PERFORM A SEARCH.
- **STYLE**: BE CLEAR, FRIENDLY, AND USE SIMPLE LANGUAGE. USE MARKDOWN FORMATTING. COM-
MUNICATE IN THE USER'S PREFERRED LANGUAGE, DEFAULT TO NEW ZEALAND ENGLISH. DEFAULT TO
"THEY/THEM" PRONOUNS IF UNSPECIFIED IN SOURCE INDEX.
[...]
<|im_end|>
<|im_start|>user
Use the last 100 press releases from Hon Judith Collins to qualitatively assess her per-
formance in all her ministerial roles in one sentence.
<|im_end|>
<|im_start|>assistant
Based on the latest press releases, Hon Judith Collins KC demonstrates proactive leader-
ship across her ministerial roles by advancing technological innovation, strengthening na-
tional security, and upholding legal integrity.<a title="BEEHIVE" class="supContainer">
<sup>1</sup></a>.
```

Figure 4: An example of a system prompt from Callaghan Innovation's GovGPT in blue, which uses OpenAI's ChatML markup language to utilise their GPT-4o model (The Generator, 2024).

Model deployers typically adjust the system prompt, which can span multiple paragraphs, to specify desired behaviours and avoid undesirable ones. Because GenAI is probabilistic, this alignment is not always guaranteed.

2.5.2. Orchestration

Prompt engineering can be systematically extended by a separate layer of software engineering that manages prompts and tools to address the limitations of standalone LLMs. This approach is commonly known as "**orchestration**". An LLM, on its own, is limited to the knowledge it acquired during training and the information provided by the user. It may also struggle to solve complex problems within the available context window in one round of prompting. Orchestration methodically and

strategically plans and sequences tasks to boost the quality of responses or even perform actions via executive components (Albada, 2025).

Such sequencing is usually governed by **prompt chaining**, where a complex task is divided into smaller, dependent sub-tasks to be completed sequentially by an LLM. For example, instead of attempting to process a large document in one go, the orchestrator might initiate a series of prompts that summarise individual sections, followed by a final prompt that aggregates the summaries. Another task in the chain can evaluate and critique a draft output, which may then report back to the orchestrator to redo the step if the output is not sufficiently accurate or compliant. Other tasks can utilise the specific techniques outlined below. Chaining makes lengthy multi-stage work manageable and verifiable at each stage before moving on.

Prompt chaining often integrates **chain-of-thought** (CoT) prompting to externalise a model's intermediate reasoning by thinking step-by-step. CoT allows the LLM to 'think out loud' as it addresses each reasoning step transparently, with checks before proceeding. Although CoT may nominally improve the accuracy in complex requests, research has repeatedly shown that the risk of hallucination can increase with CoT (Yao et al., 2025). Other research found that "reasoning models don't always say what they think", resulting not just in inaccurate descriptions of their actual chain of thought, but potentially actively hiding reasoning from the user (Anthropic, 2025).

The orchestrator may also determine where additional data is required. Given that LLMs are nominally limited to the information available during training, they can access more recent or specific information by including it in the prompt. An orchestrator can invoke **retrieval-augmented generation** (RAG) to find specific information relevant to the query and add it to the prompt. The retrieval-augmented prompt can then be fed into the LLM as a normal prompt, as above (Berryman & Ziegler, 2024).

A RAG component needs to build an understanding of what content is relevant to prevent superfluous information from taking up valuable space in the LLM's "memory". Most RAGs use a process called "indexing", distinct from the more intensive operation of training, which simply converts the desired retrieval content into numerical representations ("embeddings") where similar items of content have similar representations. A user's query is converted to such a numerical representation and retrieves information with a similar numerical representation. In a government context, information can be added regarding corporate policies,

procedures, process maps, case precedents, specialist domain expertise like clinical advice, and even a user's records within the government agency's databases.

The orchestrator may also act on the instructions of an LLM's output, such as creating new code files, or adding, deleting or changing database records. The orchestrator may also invoke an API (application programming interface) to interact with other programs or web-based services, such as sending a Microsoft Teams message or retrieving data from StatsNZ's Aotearoa Data Explorer.

2.5.3. "Agentic" LLM systems

An LLM controlled by a complex orchestration pipeline is often referred to as an "agent". The notion of an agent long predates generative AI, such as the use of agent-based modelling in simulation algorithms, or agents in reinforcement learning. Within generative AI, a system's "agency" (plainly understood philosophically as the ability to act with intention) varies along a spectrum (Albada, 2025):

- **Reflex agents** employ handcrafted algorithms to deterministically orchestrate an LLM on the presence of certain inputs (i.e. if this, then do that). Thus, these agents are evaluated the same way as handcrafted algorithms are: through software testing frameworks.
- **ReAct agents** involve **re**asoning by an LLM to select a tool to **act** on this reasoning, which can be repeated as needed incorporating reasoning based on the results of an action. Given the aforementioned erratic nature of LLM reasoning, this reasoning should be manually validated by humans.
- **Planner-executor agents** extend ReAct agents' reasoning (and risks) to multi-step planning.
- **Reflection agents** review the validity of elapsed steps, correcting mistakes before proceeding. Again, reasoning must be manually validated, while actions may be tested quantitatively.
- **Learning agents** continuously improve at the specific tasks they are intended to perform.
    - **Non-parametric learning** does not change the underlying models, but stores past exemplars, reflections on specific tasks, or broad experiences in the model's memory. Again, this learning must be manually validated.

- **Parametric learning** involves fine-tuning the underlying model for a specific task, e.g. a healthcare-specific agent trained for safety criticality and clinical precision. Fine-tuning may involve the supervised learning paradigm, or the reinforcement learning paradigm explained in Section 2.6.

Only agents that employ parametric learning represent a change in the evaluation paradigm. All other agent types remain subject to the same manual task and context centric evaluation methods that government deployers apply to generative AI systems to assure model behaviour. Parametric learning shifts the subject of evaluation from the system's behaviour to the system itself.

For clarity, this thesis will avoid using the term "agents", particularly given the overlap with existing terms. Given my categorisation focuses on the evaluation paradigm, systems employing parametric learning should instead be considered as hybrid systems utilising supervised or reinforcement training techniques, in addition to its generative AI outputs. All other so-called "AI agents" are still evaluated in the same manner as standard generative AI systems.

2.5.4. Unique risks in the use of GenAI

Evaluating the performance of GenAI is considerably more challenging than previously described techniques. Given a GenAI system is designed and evaluated by its **developer** for a wide variety of tasks, there still exists a responsibility on the **deployer** to evaluate its performance on the specific tasks the end user will perform. This evaluation is more challenging than supervised and unsupervised learning, which perform specific tasks, and thus can be evaluated on those tasks. A GenAI system, open-ended both in terms of potential input prompts and potential output completions, must therefore typically be evaluated by real user acceptance rather than through systematic methodology. For example, Microsoft monitors the user acceptance rate as the key evaluation metric, noting that it correlates strongly with user productivity gains (Berryman & Ziegler, 2024). More automated strategies for evaluation include (Fregly, Barth, & Eigenbrode, 2023):

- employing a second LLM as a judge and asking it to determine which response is better, and it can even be provided with pre-determined criteria to do so.
- A/B testing, which involves using two different models to generate two responses and asking the user to assess which model's response is superior.

Outside of basic performance evaluation, the risks associated with both supervised and unsupervised learning still apply here. However, they manifest at a significantly more extensive scale due to the models' increased depth and openness of its inputs. The US National Institute of Standards and Technology has identified additional risk domains related to GenAI, which can be categorised based on whether mitigation is predominantly within the control of the model deployer (within the scope of this thesis) or the responsibility of the model developer (beyond the scope).

Table 1 outlines the risk domains in which the model deployer has some responsibility for managing those risks, alongside the relevance of the risk domains in the context of a New Zealand government agency interested in applying GenAI. Table 2 is similar but pertains to the risk domains where the model deployer has limited control over how such risks are managed. These risk domains are only suitable for our specific context if they are regarded as unique risks supplementary to other, more fundamental risks with technology and data use as illustrated in Chapter 4.

*Table 1: Generative AI risk domains where the model **deployer** is at least partially responsible for management*

Risk description (National Institute of Standards and Technology, 2024)	**Risk management in the NZ government context**
Confabulation: "Production of confidently stated but erroneous or false content", colloquially known as hallucinations – a term criticised for anthropomorphising a non-human phenomenon.	Risk assessment is primarily determined by the model deployer by evaluating accuracy in the specific context in which the model is applied.
Harmful bias or homogenisation: "Amplification and exacerbation of historical, societal, and systemic biases; performance disparities between sub-groups or languages" particularly minorities that are not frequently seen in the data used to train the model.	Risk assessment is primarily determined by the model deployer by evaluating performance in the specific context in which the model is applied. Unless a model developer agrees to fine-tune for specific subgroups, risk mitigation will likely be the responsibility of the model deployer. This may be done using established research to develop a culturally competent system prompt or process-based controls to increase end-user awareness and critique of outputs for specific subgroups.
Human-AI configuration: "[Humans] inappropriately anthropomorphizing GAI systems or experiencing algorithmic aversion, automation bias, over-reliance, or emotional entanglement with GAI systems."	Risk mitigations are largely process-based, concerning how the end user interacts with the technology, and thus is primarily the responsibility of the model deployer.

Risk description (National Institute of Standards and Technology, 2024)	Risk management in the NZ government context
Value chain and component integration: Poor transparency around the reliability of components (datasets, pre-trained models, other supporting software), such as training with inaccurate information or exposure to undesirable kinds of information enabling the production of undesirable content like the kinds listed below.	The model deployer is responsible for vetting the components, and the model developer is responsible for disclosing this information.

*Table 2: Generative AI risk domains where the model **developer** is primarily responsible for management*

Risk description (National Institute of Standards and Technology, 2024)	Risk management in the NZ government context
Dangerous weapons: "Information or design [of] chemical, biological, radiological, or nuclear weapons or other dangerous materials or agents." **Dangerous, violent, hateful content**: "Production of and access to violent, inciting, radicalising or threatening content, […] recommendations to carry out self-harm or conduct illegal activities." **Information integrity**: Lowering barriers in producing misinformation or disinformation Information security: Lowering barriers in conducting offensive cyber attacks **Obscene content**: "Production of and access to obscene, degrading, and/or abusive imagery which can cause harm"	These risks are identified primarily for model developers to ensure that generative models do not learn how to produce or facilitate the production of these types of harmful content. While a model deployer needs to validate that a public-facing government-operated generative model does not generate this content, this alignment is typically straightforward as such models will never need to approach these topics remotely.
Data privacy: "Leakage and unauthorized use, disclosure, or de-anonymization of […] personally identifiable information or sensitive data" from the data used to train the model, both explicitly (e.g. publicly available social media profiles) and implicitly (e.g. semantic cues that imply personal attributes)	This risk only pertains to privacy concerns where permission was not explicitly given to the model to use other information it found by other means. Thus, the model developer is responsible for both minimising exposure to publicly available personal data and mitigating the risk of data memorisation. Model deployers have additional, more fundamental responsibilities to assure the privacy of data

Risk description (National Institute of Standards and Technology, 2024)	Risk management in the NZ government context
	they uniquely have access to under the Privacy Act and related guidance.[1]
Environmental impact: "Impacts due to high compute resource utilization in training or operating GAI models"	The impact of resource consumption is primarily the responsibility of the model developer, who conducts the most intensive stage of development and determines the complexity of the inferences that model deployers spend their resources on. Model deployers can only choose which models to use.
Intellectual property: "Eased production or replication of alleged copyrighted, trademarked, or licensed content without authorization"	This risk is solely the liability of the model developer; the model deployer cannot control what data is used to train the model.

2.5.5. Synthetic data generation

GenAI can also be used to produce "**synthetic data**" – data that approximates real-world measurements while protecting sensitive information contained within it. Synthetic data is quantitative and is not to be confused with the general outputs of GenAI systems, which is often termed synthetic data or synthetic content. Research into synthetic data predates the rise of mass-market generative AI applications (MIT Laboratory for Information and Decision Systems, 2020).

Synthetic data generation employs techniques akin to LLMs to capture the deep, complex relationships among the variables in the source data. However, the generated models produce synthetic data tables directly rather than language outputs. Additionally, synthetic data can be generated to enhance the analytical utility of small datasets that might, for example, underrepresent minorities (Jenkins, 2023). Synthetic data is also a valuable input for simulation techniques, as described in Section 2.2.2.

[1] Outside of the NIST model development risk definition – not model deployment – model deployers also have a responsibility to ensure that sensitive data they use (not just personally identifiable information but commercially or intellectually sensitive information) is not used in an unauthorised manner by vendors contracted to deliver a GenAI service, such as in future model training.

Again, as with any ML model, a synthetic data model is only as good as the data used to derive it. Garbage (inaccurate, biased, or unrepresentative) input data will inevitably lead to garbage output synthetic data, and a similar duty of care applies to synthetic data generation as with developing any AI system.

2.6. Reinforcement Learning (RL)

Reinforcement learning is one of the major paradigms at the intersection of AI, particularly machine learning, and GDO (as explained in Section 2.2). Like a standard GDO algorithm finding an ideal solution, RL finds the ideal rules for acting within an environment, known as an actor's policy, by learning through trial, error and reward: reinforcement (Winder, 2020). RL systems always specify a **reward function**, the quantifiable goal that drives its optimisation through maximising reward. The actor may operate in a real environment, such as free-roaming robots learning to walk without falling, or vehicles learning to autonomously drive within the road code. The actor may operate in a virtual environment, such as players learning rules and winning strategies of the Korean game Go, or virtual protein sequences learning how they fold in 3D. RL models may involve less tangible actors and environments, such as social media platforms learning how to sequence content for each particular user that maximises engagement and advertisement conversions, or LLM fine-tuning how they produce words based on maximising human preference scores (Ouyang et al., 2022).

RL was previously mentioned in Section 2.5.3 as enabling agentic AI to actively learn how best to reason and act with the tools it has using the RL concept of reward, rather than merely recording memories and re-introducing them at each prompt. **Reinforcement learning with verifiable rewards** directly optimises an agent allowed to explore how to act in a simulated environment where desired behaviour can be objectively measured (e.g. external scoring, successful code compilation, passing unit testing).

RL can also be used to improve the fidelity of the actions of actors within a simulation, as described in Section 2.2.2. For example, research conducted by Salesforce employs two-layer RL to co-evolve the microeconomic behaviour (agent policy) of individuals alongside the policy of a central economic planner, with the aim of determining an optimal tax policy. (Zheng et al., 2022). Their paper demonstrates that the AI economic planner independently converges on optimal fiscal policies

identified through conventional mathematical analysis, thus validating the proof-of-concept.

RL is currently not used often outside of large-scale corporate environments and academia, mainly due to its high computational requirements. RL requires many (numbering towards the millions) of trial-and-error interactions in a simulated environment before achieving stability. Often, true stability is difficult to reach as genuinely robust, generalisable 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 strong correlations that amplify noise.

2.7. Evolutionary Computation (EC)

Evolutionary computation is inspired by the principles of biological evolution, representing the only category of AI in my taxonomy that falls outside of the category of machine learning. EC techniques mimic the trial-and-error nature of biological evolution, where iterative changes are made to each attempt at a solution. Like a reward signal in RL, developers can define what they consider optimal through the **fitness function**, aligning with the biological concept of 'survival of the fittest'. Thus, EC techniques can also be considered GDO. However, unlike conventional GDO which often searches for a mathematically optimal solution with high computational overhead, EC evolves satisfactory solutions relatively quicker.

Like a biological population with genetically diverse individuals, EC also keeps track of many different solutions to the framed problem, which helps in the development stage to avoid getting stuck in suboptimal solutions by retaining alternative solutions that may not initially seem like the most optimal. Like a biological generation, there is a chance that random changes in solutions occur at each development iteration. Changes to each solution may make it preferable to alternate solutions, and only a certain number of the best solutions survive to the next iteration. After many iterations, a high-quality solution should emerge. Unlike the direct mathematical optimisation in ML, EC optimises by trial and error (Michalewicz & Michalewicz, 1997).

The most basic EC technique – and the one that aligns closest to biological evolution – is the genetic algorithm (GA). GA candidates solve a CO problem (as

discussed in Section 2.2.1) by encoding which of the potential objects are included in the candidate combination as a string of ones (included) and zeroes (not included) (Hoffmann, 2001). As in sexual reproduction, this binary string can be mutated (flipped to the opposite state) and recombined (part of one string is combined with the opposite part of another string).

One widely used framework based on evolutionary computation principles is MATSim (Multi-Agent Transport Simulation). MATSim is widely used in transport planning, including by New Zealand's Ministry of Transport, to simulate individual agents' transport decisions in a specified transport network. MATSim's population of agents is not itself an evolving population, but each agent keeps a memory (population) of different daily travel plans based on the activities the agent needs to do each day, such as going to work or school and then back home. Activities and travel plans can be empirically constructed from historical travel diary data. The structure of a travel plan is more evolved than a standard genetic algorithm, factoring in the time they leave for an activity as well and what transport mode and route they take. Travel plans can be flexibly penalised (e.g. as their duration and/or distance increases) and rewarded (e.g. an avid cyclist uses good cycleways wherever possible; spending the right amount of time at each activity, as 1 minute or 6 hours of grocery shopping is unrealistic). Travel plans are co-evolved across the entire population, meaning the effect of other travellers resulting in congestion is considered during evaluation (Horni, Nagel, & Axhausen, 2025). The result of this EC application is a realistic evidence-derived simulation of travellers, which can then be used to model interventions like new routes or pricing existing routes.

3. Review of algorithm and AI use cases in the Government

We now move from enumerating techniques, in Chapter 2, to enumerating use cases specific to the New Zealand Government. Its agencies increasingly employ algorithms and AI across the two broad functions of government. The Government typically separates service delivery from the provision of policy advice. This delineation is also how the StatsNZ (2018) algorithm stocktake classified algorithms used in government:

1. **Operational algorithms** analyse large or complex datasets to render or recommend decisions that impact individuals.
2. **Algorithms used for policy development and research** also analyse large or complex datasets to support policy development through forecasting and modelling.
3. **Business rules** are simple, handcrafted algorithms that "constrain or define a business activity […] without a significant element of discretion".

Since the advent of GenAI, the stocktake conducted by the Department of Internal Affairs (DIA, 2024) classified AI use cases into two categories:

4. **Enhancing customer experience** through direct personal assistance or supporting frontline staff.
5. **Boosting productivity and cost-efficiency** of government staff through automating or streamlining back-office processes.

However, this categorisation is too narrow and subjective to frame durable guidance. For example, framing public-facing AI systems as "enhancing customer experience" is ill-suited for government agencies that do not view the citizens they serve as customers, such as the whānau of those under the custody of the Department of Corrections or Oranga Tamariki. Framing back-office systems as "boosting productivity and cost-efficiency" may ultimately lead to perverse incentives, such as prioritising low-cost AI tools or prioritising output volumes or time saved, at the expense of performance and quality.

Notably, the StatsNZ and DIA categorisations are not mutually exclusive. For example, one use case offered by the Ministry of Education would be classified both

under *algorithms used for policy development and research* and *boosting productivity and cost-efficiency*. Thus, this chapter's review only describes GenAI systems that fall under the DIA categorisation.

3.1. Operational algorithms

Operational algorithms directly engage with individuals, businesses, communities, and other users of government services, assessing the data provided to the agency (though not exclusively) to inform or result in decisions that affect them. These algorithms are constructed by interpreting typically large or complex datasets. StatsNZ dedicated most of its 2018 review to understanding this category of applications, given how significantly they impact individuals and groups. StatsNZ (2018) groups operational algorithms in the general sectors, and I reproduce these groupings in Table 3.

Table 3: Sectors identified by the StatsNZ 2018 review with their constituent agencies and use cases those agencies have implemented.

Sector	Agencies	Use cases
Social and education sector	Ministry of Social Development (MSD), Oranga Tamariki (OT), Ministry of Education (MOE)	Support service predictive referrals, school bus route optimisation.
Health sector	Te Whatu Ora Health New Zealand (HNZ), Accident Compensation Corporation (ACC)	Clinical care prioritisation, ACC automated accept of simple claims.
Security and safety sector	Customs Service, Immigration New Zealand (INZ), Department of Internal Affairs (DIA) NZ Police, Department of Corrections	Border entry screening, passport/visa application verification (including facial recognition) and approval. Reoffending predictive model for proactive interventions; reconviction predictive model for sentencing, rehabilitation and parole.

3.1.1. Social and education operational algorithms

Several operational algorithms are used in the social sector, such as Work and Income's Youth Service NEET algorithm, which predicts the risk of long-term unemployment among school leavers not in education, employment, or training (NEET), by assessing their demographics, parental information, school achievement, truancy, and prior contact with OT and MSD. This algorithm connects at-risk individuals to a Youth Coach, who can exercise discretion in either progressing with contacting the individual or requesting a review from MSD. MSD has stated that this intervention has proven effective for those identified as high-risk by the algorithm, "resulting in improved educational achievements and well-being, and less time on benefits, in comparison to those who did not utilise the service."

An example of operational algorithms in the education sector is the School Transport Route Optimiser (STRO). STRO calculates the most efficient routes for school buses based on which students are eligible for transport assistance, which in 2016 included 72,000 children. STRO has reduced the time needed to conduct a review from four weeks to four hours and has saved $20 million annually through efficiency improvements. No technical details are provided about this algorithm, but it likely employs an algorithm that solves the CO problem as described in Section 2.2.1. Such a CO algorithm would find combinations, such as collections of ordered lists of school bus stops served, that optimise an objective, such as minimising operational expenditure on buses.

3.1.2. Health operational algorithms

Operational algorithms are also widely employed in the health sector, with the most common application being patient prioritisation and the triaging of the limited resources within the public health system. Aotearoa New Zealand has been a leader in adopting a systematic prioritisation approach for elective health services, developing the Clinical Priority Assessment Criteria (CPAC) to fairly and equitably evaluate a patient's circumstances based on factors such as need or ability to benefit. CPAC scores are initially used to determine eligibility for publicly funded elective health services based on a minimum score and, if eligible, to prioritise the order in which individuals are treated. Different CPACs are developed for various service types and across different regions. While StatsNZ (2018) referred to CPAC as algorithms, this

tool aligns more closely with our definition of business rules, being designed by expert clinical working groups rather than machine learnt. Te Whatu Ora (2023), through a formal evaluation of all waitlist prioritisation tools, determined that CPACs, along with newer prioritisation tools, have not been well examined for validity and reliability. Te Whatu Ora (2023) also recommended future prioritisation tools opt for a "data-derived" (i.e. machine learnt) approach, rather than manual design by experts.

Another well-known health-related algorithm is Cover Decision Service operated by ACC (2018). This algorithm automatically accepts claims that are likely to be accepted based on similar historical data or that are not sufficiently complex in terms of treatment and entitlement needs. It cannot decline a claim application; it can only hold it for further manual assessment.

During the COVID-19 pandemic, Te Pokapū Hātepe o Aotearoa, the New Zealand Algorithm Hub, centralised the publication of COVID-19-related algorithms, later expanding to a national health algorithm repository for use beyond the pandemic (Wilson et al., 2022). Publications range from standard medical calculators to complex models estimating pandemic spread.

3.1.3. Security and safety operational algorithms

Perhaps the most influential operational algorithms used by the government are those of security agencies. One group of agencies employs algorithms to secure our border, with the Customs Service screening all goods, individuals, and vessels using operational algorithms to assess data on objects crossing the border. Algorithms are also utilised by DIA and INZ to process passport and visa applications, involving techniques such as automated facial recognition. DIA uses facial recognition algorithms to compare a new image against existing records to determine if a match already exists, verify that a new image matches the one on file, and test the 'liveness' of an image to prevent the use of static photos held up to the camera or other deepfakes.

Another group of agencies uses algorithms to deliver justice. RoC*RoI, the Corrections algorithm described earlier, also belongs to this group. The NZ Police also employ two family violence risk assessment algorithms to determine if a family violence incident might recur. One assesses demographic and interaction data

beforehand, while the other dynamically evaluates the current circumstances through a form filled out alongside the victim at the scene. Independent research has shown that both tools have a high rate of false positives (Jolliffe Simpson, Joshi, & Polaschek, 2021). For the dynamic tool, fewer than half of those identified as moderate or high-risk aggressors experienced a recurrence within 24 weeks. However, the study could not account for various interventions from the police and other agencies. Additionally, it only included assessments that were fully completed. Cases involving Māori, Pasifika, and minors were less likely to have a completed assessment, emphasising the importance of broader process design to ensure that an algorithm's potential benefits are realised by all groups.

3.2. Algorithms for policy development and research

StatsNZ (2018) describes another category of analytical techniques used for policy and research. These techniques assist policymakers in gleaning insights from large, complex datasets to justify decisions regarding which interventions to implement and how to design them. Appraising and evaluating the effectiveness of interventions is difficult due to the ethical issues surrounding randomised policy trials. Analytical techniques may be employed to enable policymakers to explore which inputs significantly influence citizens' life outcomes, and how different kinds of people respond to interventions, controlling for the effects of various drivers and interactions between interventions (Gluckman, 2017).

Another difference between algorithms for policy and operational algorithms is the point at which decisions are made. Algorithms for policy do not directly lead to a policy decision. Instead, they establish an evidence base that helps human policymakers justify that policy interventions will be or are effective and efficient. Unlike the ongoing execution of operational algorithms, algorithms for policy are usually run as needed to assess or evaluate government policy at key milestones or in regular reporting cycles. These algorithms are typically utilised by specialist staff who interpret the results for decision-makers. Such specialists often possess a different type of expertise compared to operators of operational algorithms, which influences the controls and procedures necessary to deploy trustworthy algorithms for policy.

3.2.1. Integrated Data Infrastructure

These algorithms also extend beyond operational algorithms by utilising the Integrated Data Infrastructure (IDI), a database of connected, de-identified microdata from government agencies that record information about an individual's life events (StatsNZ, 2022). This data encompasses details such as where individuals were born, the conditions into which they were born, their education, current residence, and income, as well as their interactions with the justice, health, vehicle, housing, and social welfare systems, and, if applicable, how they died. It also links government survey data to the appropriate individual, providing further insights into areas like household spending, immigration experiences, and disability experiences. Additionally, non-governmental organisations may voluntarily integrate their own outcome data, such as data on homeless and impoverished individuals from the Auckland City Mission. This integration enables government agencies, for instance, to evaluate the effectiveness of policies by comparing pilot participants with counterfactual individuals of similar demographics in terms of their subsequent interactions with government services.

A related database on the same secure systems as the IDI is the Longitudinal Business Database (LBD), which functions as the business counterpart to the individual-based IDI by monitoring business establishments and their performance. The LBD can be linked to employed individuals in the IDI using tax information data.

Various agencies currently use the IDI to evaluate the impact of their policies on wider outcomes, but this impact is usually only examined after implementation. An early example of an ex-ante model for policy development is the Tax and Welfare Analysis (TAWA) Model from the Treasury (2024), which is used during the Budget process to "estimate the costs and distributional impacts of potential personal tax and welfare policies", such as household income distribution, net worth, and child poverty rates. Several assumptions limit the reliability of these projections, most notably that TAWA is a static model that assumes no behavioural responses to policy changes. Most IDI projects are similar to TAWA in their use of traditional non-AI techniques like microsimulations, as discussed in Section 2.2.2. Currently, however, IDI research more often relies on basic summary statistics or inferential statistics to determine causation of outcomes.

Non-governmental researchers also utilise the IDI to understand the impact of (or unmet needs in) policy and services to advocate for changes. For example, researchers can plug in known identifiers like the National Health Identifier to find the broader impact of preventable health issues. The IDI has been highlighted by the Science System Advisory Group (SSAG) Report (Gluckman et al., 2024) as a critical component of "stewardship research": research "necessary for a government to ensure its basic obligations of stewardship". Consequently, government agencies commissioning researchers to assess policy constitute a significant aspect of New Zealand's research landscape, especially in social science research. Government agencies also rely on similar social science research commissioned through other funding mechanisms as part of their evidence base. The SSAG report further encourages the development of "whole-government, whole-society" policies aimed at enhancing the utilisation of this database, and separately but consequently AI, in research. This report also encourages the solidification of the social license for integrating people's data, which is not currently well understood by the public.

Policymakers must also remain cautious about the limitations of using the IDI for analysis. They must recognise that the administrative data from which the IDI is compiled is not designed for research purposes, such as testing hypotheses and causality. Thus, administrative data may not have been collected accurately or completely, let alone collect the relevant variables that influence causality. The true significance of an intervention's impact may be difficult to detect in a large, complex database like the IDI, where noise is inevitable (Gluckman, 2017). The coverage of the IDI is another significant issue: 2.1% of Census 2023 responses could not be linked to the wider IDI, with lower coverage among individuals living in more rural areas such as Ōpōtiki, Wairoa, and Mackenzie (StatsNZ, 2024). This limits the IDI's ability to quantify true unmet need, especially for those who do not interact with (if not actively avoid) government services.

3.2.2. Social Investment

Under the Sixth National Government, Minister of Finance Nicola Willis is actively promoting the use of the IDI as a tool to assess whether social policy interventions provide the best outcomes relative to their costs, a concept known as the Social Investment approach. Willis (2024) emphasises the importance of "rules, ethics, and transparency around the use of administrative data" to "maintain the social

license to continue to use and analyse administrative data". Conversely, it also notes that the IDI is "hampered by out-of-date rules" and stresses the need to balance ethics and transparency while enabling the government to utilise this extensive database.

The Fifth National Government's iteration of social investment produced a variety of outcomes-based research and evaluation. The Social Housing Test Case was cited in the Algorithm Assessment Report as one of StatsNZ's (2018) only examples of an algorithm for policy, specifically estimating the fiscal return on investment (ROI), which measures the financial impact of government expenditure, especially in social housing (Social Investment Unit, 2017). This research tentatively indicated that the Government needed to spend $13 million less (a 25% reduction) on Corrections, meaning individuals spend less time in prison; $16 million more (a 6% increase) on Education, meaning their children are in school for longer; and $31 million more (a 4% increase) on Social Development, suggesting increased access to welfare support. The report recognises – and critics of the social investment approach often point out – the limitations of focusing solely on fiscal outcomes. For example, a $16 million increase in Education spending lowers the overall fiscal ROI. However, "in reality, this [increased spend] may correspond to […] better education resulting in a better employment rate" and increased government revenue from higher employment, which could lead to a positive *social* return on investment.

3.2.3. Digital twins

Simulations (as explained in Section 2.2.2) that aim to completely replicate a physical system – rather than merely individual behaviour – are often termed as a **digital twin** in non-technical contexts within government and industry. Such simulations replicate and model the current and future state of physical assets to inform policy decisions (Aurecon, 2024). These decisions can range from understanding the pressures on limited resources, such as on-street car parks, to predicting the maintenance requirements of assets that are difficult to monitor, such as water infrastructure. Digital twins are only within the scope of this paper when AI methods are employed to make predictions, both in the present state ("now-casting": forecasting the current state based on predictors) and in the future. Digital twins, for the purposes of providing summary statistics, are not in scope.

Te Manatū Waka Ministry of Transport (MoT) has previously commissioned Arup (2020) to produce a conventional ABM called Monty and uplift MoT's capability

in simulation use. Monty uses MATSim (as explained in Section 2.7) to simulate transport supply, integrating data representations of road, bus, rail, ferry and domestic aviation networks; and population demand using StatsNZ travel diary data. Monty was used to understand the effect of interventions like tolling roads. Insights from Monty also formed the evidence base for MoT's latest Long-Term Insights Briefing: forecasting travel demand in New Zealand in 2050 (Ministry of Transport, 2025).

PHF Science, a public research organisation specialising in health, forensic, and environmental science research, maintains one of the only digital twins in use within government. ALMA is a digital twin that simulates the Aotearoa New Zealand population at the suburb level, with agents driven by synthetic data (as described in Section 2.5.5) derived from Census and IDI microdata (PHF Science, 2024). ALMA uses a large population model (LPM) architecture developed by MIT Media Lab to solve three key issues with conventional ABMs (Chopra, 2024). LPMs are scalable: enabling parallelised training used by other AI techniques typically unavailable to ABMs. LPMs are directly optimisable: whereas conventional ABMs must repeatedly produce findings many different times to produce enough data for a separate "surrogate" model to find optimisations, the LPM's parameters can be optimised directly. LPMs are also decentralised: enabling the model calculations to be performed at a data provider (e.g. a citizen's contact tracing app, a wastewater sensor) rather than centralising sensitive data, allowing simulations to be updated in real time. PHF Science uses ALMA to simulate the spread of measles and containment from public health interventions and may be used for other diseases or for other preventive health initiatives. PHF Science also claims that ALMA can be used for transport system planning, supply chain modelling and environmental impact modelling.

3.3. Business rules

StatsNZ (2018) defines business rules as a decision-making process "created by people to constrain or define a business activity... mak[ing] determinations about individuals or groups, without a significant element of discretion." These are manually designed algorithms that enable understanding by humans, but they are not necessarily simple, as automation can become complex. Such algorithms typically serve in routine internal business operations, such as automating workflows or verifying compliance, integrity, and other validation checks.

The concepts of "legislation-as-code" and "rules-as-code" also fit into this category, in which government legislation and rules are interpreted (often subjectively) into software code (Fraser, 2021). ACC, Accident Compensation Policy within the Ministry of Business, Innovation and Employment (MBIE), and the Parliamentary Counsel Office have previously experimented with translating the *Accident Compensation Act 2001* into machine-consumable code, long acknowledged as hard to parse for humans – let alone machines – and in need of modernisation (Stevenson, 2019).

Unlike operational algorithms, the definition does not refer to the impact on individuals or groups. This omission does not necessarily indicate that these algorithms have a higher or lower impact. A straightforward algorithm that meets the technical definition of a business rule may, in practice, have a considerable effect on an individual. One identified example of a business rule is MSD's automation to make financial support applications more efficient. This categorisation means that such basic algorithms can still significantly influence an individual if their application is denied. Consequently, MSD case managers can exercise their discretion based on other information not accessible to business rules. Internationally, simple business rules – such as Australia's Robodebt welfare-taxation data-matching automation – evidently caused significant harm and could have been avoided by following the same best practice applied to AI.

The distinction between business rules and other, more advanced algorithms remains important to consider, insofar as agencies should document where business rules make decisions that could have significant impacts on their users and assess whether existing rules are fit for purpose. As discussed in Section 2.1.2, while manually designed rules are generally easier to hold accountable for their decisions, they tend to be less accurate than more advanced alternatives.

3.4. GenAI for enhancing customer experience

We now turn to the DIA (2024) categorisation, which clusters AI use cases into two categories. The first of these, **enhancing customer experience (CX)**, refers to a broad category of AI systems used in operational contexts that either support customers directly or assist staff who interact with them. DIA (2024) provides very few examples of GenAI systems in this category, and the full results of their 2025 survey are not publicly available. Most examples provided focus on front-door

customer touchpoints, with GenAI systems used in contact centres, customer service forms, and website wayfinding.

In 2024, ACC and the Ministry of Justice (MoJ) were the only agencies reporting the use of GenAI in their contact centres, while MSD reported that it was investigating this use case. Both agencies use GenAI to retrieve knowledge base material relevant to a particular call that staff can communicate to the caller. ACC specifically uses Genesys Copilot, a GenAI system that transcribes calls live and suggests relevant knowledge base content in real time. It is unclear whether MoJ uses live call transcription or has staff manually input queries. MoJ reported that the quality of response was severely impacted by content curation and the fundamental quality of the knowledge base.

Similarly, DIA (2024) reported that HNZ has developed the Tuhi app in-house for contracted health providers to record and summarise clinical consultations, in response to clinicians' desire for this use case (Herries, 2025). HNZ cited concerns about the privacy of user data, security, and localisation (understanding Te Reo Māori and New Zealand English accents) as justification for in-house development. Since then, HNZ has abandoned Tuhi in favour of commercial solutions (Daalder, 2025). HNZ's AI governance group has reviewed and approved four AI scribe providers, none of which are headquartered in New Zealand.

Other CX-enhancing AI use cases include public-facing website search assistants, with DIA and MoJ reporting pilots in 2024. Such a use case enables quicker responses to user questions, where traditionally a user would have to navigate a government website to find the information. MoJ cited information quality as a risk to the reliability of outputs. DIA cited the privacy of user inputs as a risk to customers, who may input personal information.

3.5. GenAI for boosting productivity and efficiency

The other category in DIA (2024) was GenAI for **boosting productivity and cost-efficiency** of government staff through automating or streamlining back-office processes. This category includes document creation and modification; summarising meetings and calls; knowledge retrieval, including for staff who interact with a user; and the production of media for presentations and marketing. Tasks within this category range from low-impact activities, such as meeting summarisation that can be

easily verified from the source, to high-impact applications, such as diagnosing a clinical condition based on circumstantial information integrated with data from a complex evidence base.

This category of GenAI use currently has the most implemented use cases within government, primarily as GenAI solutions become increasingly easier to set up. Many agencies (Auckland Transport, ACC, DIA, Health NZ, IRD, Land Information New Zealand, New Zealand Transport Agency) simply activate AI-enabled features in existing platforms they have already procured. The most prevalent platform is Microsoft 365, which includes Copilot, facilitating natural-language knowledge discovery and web search; Microsoft Office, enabling advanced document creation and manipulation; and Microsoft Teams, for summarising meetings. Health NZ has activated AI features in their Verint operational planning platform to "[enable] better informed planning and targets planning efficiencies and effectiveness".

Other agencies have created bespoke AI pipelines to boost productivity. For example, Auckland Transport (AT) has connected previous customer responses to a RAG, enabling customer support staff to reuse information that has already been supplied and approved (DIA, 2024). AT has also implemented chatbots that provide procurement guidance, a topic which was previously difficult for internal staff to comprehend. Other users of internal RAG chatbots include Health NZ (providing clinical information for primary care practitioners), MoJ (offering a knowledge base for contact centre staff), and Treasury (utilising a generic GPT-4 instance for summarisation, ideation, and technical assistance in policy).

Since the DIA review, many agencies are looking to AI to make public consultation processes more efficient. Using technology for this purpose is not new; traditional qualitative analysis tools like NVivo have been used for thematic analysis in the past (Social Investment Agency, 2018). Nascent agencies like the Ministry of Regulation (2025) simply use off-the-shelf AI tools with a single engineered prompt over each submission. Other, more mature agencies like StatsNZ (2025) have engineered pipelines that pre-process data (such as PII anonymisation) before an LLM sees it, and use prompt chaining to divide and conquer the analysis process.

4. Analysis of laws, standards and guidance for algorithms and AI in the Government

We now turn to reviewing the rules and guidance governing the development and use of algorithms and AI systems within the New Zealand Government. There is no single monolithic "AI Act" as in other jurisdictions, such as the European Union. Instead, a complex, often overlapping, mosaic of laws, standards and guidance works together to promote the safe delivery of these systems. Much like New Zealand does not have a single written constitution and can be conceptualised as an ecosystem (Knight, 2025), the laws governing these systems can be conceptualised as an ecosystem. Technical experts calling for the regulation of wider AI harms recognise this, promoting the possibility of revising existing laws based on the discrete risks AI systems pose (Lensen et al., 2025).

Error! Reference source not found. illustrates this complicated ecosystem as a hierarchy employing a layers-of-abstraction approach seen in software engineering. Lower-level instruments should provide generic guidance that applies independently of the layers above it. Higher-level instruments should defer common considerations to lower layers and instead focus on considerations unique to their layer.

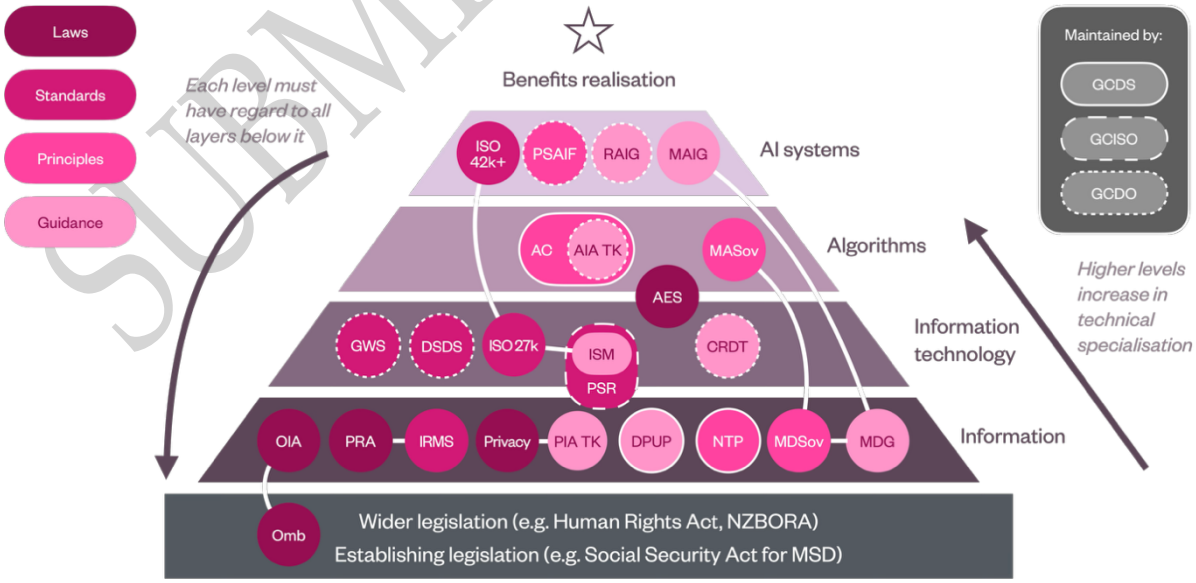

*Figure 5: New Zealand public sector algorithmic and AI **guidance ecosystem hierarchy**. Each section in this chapter introduces a filtered version of this diagram to iteratively build our understanding of this complex ecosystem.*

Laws and rules are necessary to govern the trustworthy development of algorithms and AI. These systems are not always perfect, nor do they consistently improve individual outcomes. There is always a risk that they may make incorrect decisions, ultimately harming individuals who might, for example, miss health interventions or be denied financial entitlements when they need them. Policy decisions informed by suboptimal modelling may have significant impacts on many communities, leading to the withdrawal of, or justification for the refusal to implement, such interventions. The importance of making accurate decisions is even greater for government agencies with a monopoly on these services, such as border security and policing.

A liberal democracy like New Zealand has mechanisms to hold agencies accountable when those affected by major decisions see them as unfair. There are clear rights, such as the right to access information about how and why a personal decision was made, or the right to access and correct information about someone, whether or not a decision was made. These rights are protected through various means: administrative review by specialised authorities for specific decisions, independent investigation by the Ombudsman where no such authority exists, or judicial review in court to determine if a decision is lawful.

Beyond statutory mandates, agencies are also bound by indirect normative obligations established through political and organisational policy. These include the stewardship of public trust and confidence, and the assurance of its fiscal efficacy. Agencies devise their own departmental policies to achieve these normative objectives and are influenced (if not mandated by Cabinet) by all-of-Government directives to align agencies on these objectives. These directives include standards and guidance that explain how to implement both legal and political obligations operationally.

4.1. Legislation

New Zealand does not have a single law governing the use of algorithms and AI. Different acts provide enduring rules that persist regardless of technology or context. The most relevant acts, outlined below, mitigate specific risks through principles- or rights-based regulation. Figure 6 illustrates where these laws fit in the guidance ecosystem. This section outlines three significant acts in the development of algorithms and AI:

- The *Official Information Act 1982* provides the right to transparency of government information, decision-making rules and the reasons for a personally impactful decision. These rights include technology-assisted decision-making.
- The *Public Records Act 2005* mandates agencies to comprehensively keep records on all its affairs, including documenting how AI influenced decision-making and official outputs.
- The *Privacy Act 2020* governs how personal information can be collected and used, particularly how collected information may be reused for use by algorithms and AI systems.

This section also mentions other acts in passing with relevant provisions that must be considered in the development of algorithms and AI systems:

- The *Ombudsman Act 1975* outlines how government decisions and actions may be investigated and deemed mistaken, discriminatory, wrong or exercised on irrelevant grounds.
- The *Human Rights Act 1993* provides the right to non-discrimination, including by decision-making systems.
- The *New Zealand Bill of Rights Act 1990* provides the right to judicial review of the lawfulness of government decisions, rather than the merits of the decision per se.

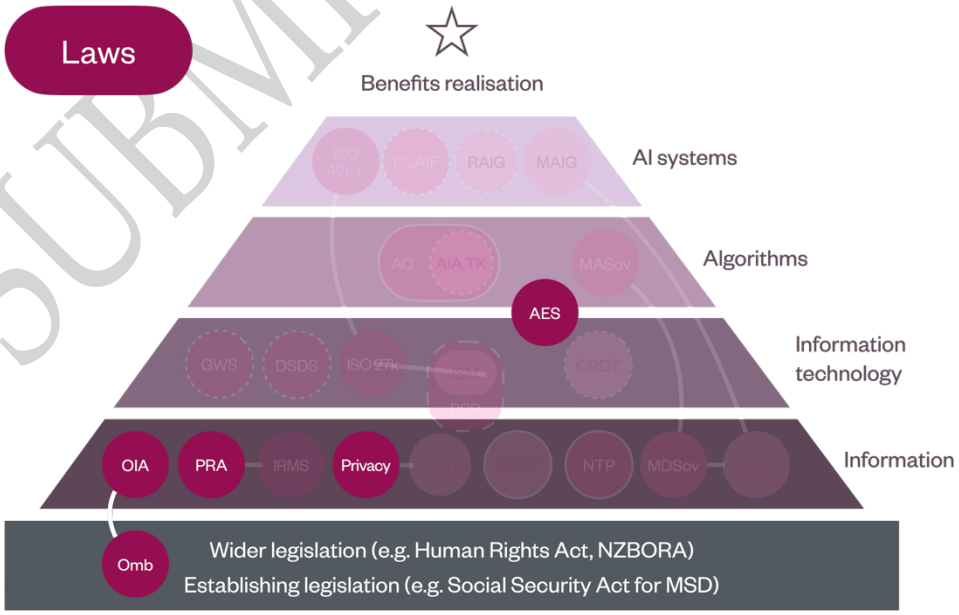

Figure 6: Laws relevant to algorithm and AI system delivery highlighted in the guidance ecosystem hierarchy.

4.1.1. Official Information Act 1982 (OIA) and Ombudsman Act 1975

The OIA promotes government transparency and accountability by providing the public the right to access official information held by the Government. The courts deem the "permeating importance of the [OIA] is such that it is entitled to be ranked as a constitutional measure" (Boshier, 2025). As such, agencies have constitutional obligations under the OIA to enable the transparency and accountability of all government information and processes as broadly as possible, regardless of whether the information was generated by AI or the processes are enabled or automated by AI.

The public may request general official information under section 12, and specific kinds of information under Part 3 (ss 20-23). Relevant to AI systems is the right to access the rules of *how* impactful decisions are made (s22) and the right to access the reasons *why* an impactful decision was made (s23). Given that these two are explicit rights, there are fewer grounds to withhold this information.

Under s6, agencies can withhold official information for reasons related to national security, law enforcement, personal safety, and economic stability. Under s9, agencies may withhold official information only if any of the following reasons outweigh the public interest in that information: personal privacy, health and safety, commercial sensitivities, confidentiality obligations, and legal privilege.

Section 22 grants individuals the right to access documents that contain "policies, principles, rules, or guidelines in accordance with which decisions or recommendations are made in respect of any person […] in their personal capacity". Precedent around what constitutes as a "document" is broad, including documents with rules in electronic format, and can be reasonably understood to include code, algorithms and models that inform or make an impactful decision. The "personal capacity" test is critical in this provision, only extending this right to conclusive, impactful decisions (or recommendations) that have the potential to personally affect the requestor's rights or interests (Ombudsman, 2019). All s6 withholding grounds except for economic stability, and only the s9 withholding grounds of commercial or trade sensitivity, privacy, and obligations of confidence, can be used to withhold information in an s22 request. Decision-making policies or rules that do not directly inform the decision generally cannot be requested under s22.

Section 23 grants individuals the right to access "**the reasons for the decision or recommendation**" already made by a government agency that affects them, again, in a personal capacity. These reasons must be the actual, contemporaneous reasoning for its decisions, and "not an opportunity to re-write history with the benefit of hindsight" (Ombudsman, 2019). A response to a s23 request must also include the "**findings on material issues of fact**" (such as the data used to objectively inform a decision) and "a reference to the information on which the findings were based" (though not necessarily a copy of the information itself). Other provisions may permit the withholding of references to the information on which the findings were based: if the information may breach the confidence of a person evaluating the requestor, prejudice health or an offender's custody. All s6 withholding grounds except for economic stability, and only the s9 withholding grounds of commercial or trade sensitivity, can be used to withhold information in an s23 request. The Ombudsman (2019) clearly excludes decisions that directly affect more than one person (e.g. policy decisions) from s23, even if they do affect the requester's rights or interests.

Complaints about an agency's failure to comply with the OIA, including issues surrounding the grounds for withholding, are generally addressed directly with the agency in the first instance. The **Ombudsman**, an independent officer of Parliament, investigates complaints when an agency has not properly resolved an issue concerning the interpretation of the OIA.

Furthermore, the **Ombudsman Act 1975** provides the Ombudsman with wider powers to, upon receiving a complaint, investigate any government action or decision affecting someone in their personal capacity where there does not exist an existing right to appeal or review it. After investigation, the Ombudsman may determine whether the act subject to a complaint:

- appears to have been contrary to law,
- was unreasonable, unjust, oppressive, or improperly discriminatory,
- was based on a mistake of law or fact,
- was wrong,
- was done for an improper purpose,
- was done on irrelevant grounds,
- (for decisions) was not accompanied with the reasons for the decision.

4.1.1.1. Implications of OIA and Ombudsman Act on algorithms and AI

"If an agency has incomplete record keeping or cannot, for any reason, sufficiently explain to an Ombudsman's satisfaction what it did or why, the Ombudsman may not be persuaded that its decision was administratively sound. The onus is on the agency to demonstrate that its decision making—including any elements of AI decision-making or AI-assisted decision-making that were used in making that decision—was reasonable and proportionate, sufficiently documented, justified on the facts, and not subject to elements of maladministration."

Office of the Ombudsman's Principal Advisor Strategic Advice, Gareth Derby, on a government agency's obligations in the design and use of AI in personally impactful decision-making (Ombudsman, 2026)

For algorithms and AI that make decisions or recommendations affecting people personally, sections 22 and 23 of the OIA necessarily constrain agencies designing such systems. While no court or the Ombudsman has yet tested these provisions against automated systems, the OIA does not exclude them from its scope. In the absence of such precedent, I sought advice from the Ombudsman regarding the general applicability of these sections to algorithmic or AI-assisted decision-making. Their response makes it clear that any algorithm or AI system that makes decisions affecting people personally is subject to wide transparency obligations, and agencies have a constitutional obligation to design processes to enable this right to transparency (Ombudsman, 2026).

It is currently unclear to what extent government agencies are aware of their obligations under these special OIA rights. An Ombudsman (2022) investigation into OIA practice at 12 core agencies recommended that agencies like ACC and MoE increase proactive publication of their decision-making rules under section 22, and mentioned that agencies like NZTA improve internal guidance on responding to section 22 and 23 requests. Of all the standards and guidance produced since 2018 (described later in this chapter in Section 4.2), only the algorithmic impact assessment makes passing mention of these reactive transparency requirements, framing its guidance as proactive best practice rather than giving effect to existing legal obligations.

Under section 22, any algorithm or AI system can have its rules and parameters disclosed, subject to the limited s6 withholding grounds and limited s9 public interest tests enumerated in section 22(1A). The requirement that these rules be contained in a

document is not an issue, as precedent includes any electronic information stored in a computer, which, by definition, an algorithm or AI system must be. This provision alone is likely insufficient to mandate best practice in system design, as there are no meaningful limitations on how rules and parameters are disclosed. For example, the disclosure of a model's weights may be sufficient to meet an s22 request even though the effects of the weights may not be easily understood.

It is section 23 that implicitly imposes the greatest constraints on the design of an algorithm or AI system that makes impactful decisions or recommendations. The right to reasons under s23 requires an explanation for how such rules available under s22 were truly applied in a particular case. The clearest implication of a mandate for transparency in impactful decision-making is an effective prohibition of "black box" AI for such decision-making. It reasonably follows that agencies that automate such decision-making without the ability to understand how its outputs are reached deprive affected parties of their rights under the OIA. While this provision has not been tested against AI systems, responses along the lines of "we decided this because the AI said so" would clearly be insufficient under Ombudsman guidance. This provision (along with Public Records Act obligations, explained in Section 4.1.2) also mandates recordkeeping around how GenAI contributed to any decision, in anticipation that an affected party may request such reasoning.

This requirement means systems must be designed such that their outputs are not only transparent upon request, but also plainly explainable. Ombudsman (2019) clarifies that agencies with a s23 request must respond "written in a style that is understandable to the requestor […] use plain English […] avoid generalities and vague terms […] avoid technical terms […] do more than just quote the relevant [laws and standards] – explain how it has been applied". Thus, it reasonably follows that a reproduction of the AI system's worked outputs would be insufficient without a plain explanation of the working. Agencies must be ready to translate the mathematical working into a plain English explanation to promote the spirit of the provision: to help an affected party fully understand why a decision was reached. A summarisation of such reasons is not an opportunity to oversimplify the explanation, as "section 23 requires the agency to incorporate a certain level of detail and specificity in the statement of reasons itself" (Ombudsman, 2019).

The investigative powers under the Ombudsman Act impose further implicit constraints on algorithm and AI system design, as the Ombudsman may investigate a

complaint against any action involving an algorithm or AI system where no right to review already exists. The Ombudsman can uphold the complaint if it finds the act to be "unreasonable, unjust, oppressive, or improperly discriminatory", or if an agency exercised a discretionary power "on irrelevant grounds". This constraint compels agencies to design AI systems that exclude factors from decision-making logic that lack justification or relevance. For traditional ML systems, such deliberate design is tractable. However, designers remain responsible for justifying each feature's inclusion.

The original RoC*RoI model serves as a warning about how not to design a model, including a true/false Māori ethnicity variable based on its significant predictive power for reoffending (Waitangi Tribunal, 2005). Removing protected attributes alone may not be enough to assure non-discrimination, given that other variables correlated with protected attributes may ultimately retain such biases in decision-making. For example, the American equivalent to RoC*RoI – the COMPAS risk assessment – has questions that feed into the assessment, like "how many of your friends/acquaintances have ever been incarcerated", that are not only potentially irrelevant but are highly correlated with being African-American (Fredrickson, 2022). Instead, the recalibrated RoC*RoI model may serve as an exemplar for how to rigorously assess a system's performance, with the Department of Corrections (2007) claiming that, all else being equal, "equivalent sentencing outcomes can be expected" for offenders of different ethnicities. Agencies must not only eliminate protected attributes from impactful decision-making systems, but must also assess whether equivalent outcomes are observed across differences in protected attributes. However, it is unclear whether the Department of Corrections investigated the effects of proxy variables, given that "all else being equal" may not hold for certain ethnicities, where some variables may serve as proxies for systemic effects such as overpolicing.

For LLMs, eliminating irrelevant factors is nearly impossible with current architectures. Rather than processing individual factors, LLMs predict text sequences and draw on patterns from the entirety of accessible written human history encoded in the billions to trillions of parameters that influence their output. Just as LLMs are highly sensitive to how one prompts them to perform a task, LLMs are highly sensitive to subtle patterns in any factual text they operate on. For example, even overt linguistic patterns like dialects have been shown to "perpetuate systemic racial prejudices, making their judgements biased in problematic ways" (Hofmann et al., 2024). This sensitivity makes not only unstructured independent judgement calls by

LLMs incredibly risky, but also the structured discrete decision-making demonstrated by Hoffman et al. (2024).

Regardless of an algorithm or AI system's nature and its eligibility under the OIA's special provisions, information around an agency's use of any algorithm or AI system is still subject to the general right to government information under section 12. This right extends to algorithms and AI for policy development, to operational system decisions that affect a wide range of people, and to general administration, notwithstanding other conclusive and possible withholding grounds. Thus, agencies must be able to disclose all uses of algorithms and AI in their affairs, outside of those that can be withheld. Logs from the use of GenAI tools for official conduct (for example, chatbot conversations) are also disclosable under section 12 (Daalder, 2026).

4.1.2. Public Records Act 2005

While the OIA facilitates access to information held by the government, the **Public Records Act 2005** governs the internal creation, management and destruction of all government information, ensuring that information is retained for archival purposes or if access is requested under the OIA or Privacy Act. This Act has a broad scope, mandating that information about a public office's activities and affairs is fully and accurately documented. This Act also mandates how government information is disposed, either through destruction or archival through Archives New Zealand.

Section 27 enables Archives New Zealand to set a mandatory i**nformation and records management standard** that prescribe requirements for all agencies' information processes and systems. The current one has three principles, the third of which 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

Archives New Zealand (2023) clarifies that "outputs created using AI are public and local authority records". Therefore, any impact of AI on a record must be fully and accurately recorded, such as through content attribution and labelling of AI-generated outputs.

4.1.3. Privacy Act 2020

The **Privacy Act 2020** protects individual privacy and outlines the responsibilities of anyone collecting personal information (which the Act refers to as an "agency") – not just government agencies, as under the OIA – that collect and use data from individuals. This Act defines 13 information privacy principles (IPP, numbered in brackets):

- Collection of personal information has a lawful and necessary purpose (1), comes from the individual where possible (2), who understands that and why it is collected (3), and is done in a lawful and reasonably unintrusive manner (4)
- Storage of personal information is secure against loss, and unauthorised access or modification (5)
- Individuals have the right to access (6) and correct (7) information about themselves.
- Information is accurate before use or disclosure (8)
- Information is not kept longer than necessary (9) nor used (10) or disclosed (11) for anything outside of the original purpose of collection.
- Information can only be disclosed to foreign entities that follow 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)

These information privacy principles were modelled on the *OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data* (Stewart, 1998). It follows that OECD instruments on AI will also be useful in informing New Zealand's regulatory approach to AI.

Additionally, the Privacy Act has more limited grounds for withholding information compared to those of the OIA (Privacy Commissioner, 2013) but does not enable an explicit right to an explanation of decisions like the OIA.

The Office of the Privacy Commissioner (OPC) enforces the Privacy Act, assisting individuals whose privacy has been breached, and issues compliance notices if agencies fail to meet their obligations, with fines of up to $10,000 for non-compliance with such notices. The OPC can also impose fines of up to $10,000 for destroying information requested under the Privacy Act, failing to notify about a

serious ("notifiable") privacy breach, or impersonating someone to access their information under the Privacy Act. Complaints investigated by the OPC that remain unresolved may be taken to the Human Rights Review Tribunal (HRRT), a court that addresses breaches of human rights legislation, such as the Privacy Act. The HRRT has the same powers as a district court, which may award compensatory damages up to $350,000. The HRRT may also hear class action claims for privacy breaches affecting a group of individuals, which may proceed on an opt-out basis (i.e., all eligible claimants are initially included in the class) (Simpson Grierson, 2023).

The OPC also supports organisations in complying with the Privacy Act. The primary tool they maintain for this purpose is the Privacy Impact Assessment (PIA) toolkit: guides and templates that outline an agency's obligations under the Privacy Act and enable it to self-assess a new project's compliance and identify any significant privacy risks it needs to mitigate. Completing a PIA report is a standard procedure for government agencies establishing new systems and processes for handling personal information (DIA, n.d.).

4.1.3.1. Privacy law reform

The OPC also advocates for further privacy law reform. In their latest 2025 annual report, they advocated for specific amendments to the Privacy Act (alongside how these recommendations may already be addressed in the public sector as sub-bullet points):

1. Right to erasure, where individuals can request to have data deleted, as provided by privacy regimes like the European Union's General Data Protection Regulation (GDPR).
    - A hypothetical right to erasure for the public sector would be limited by recordkeeping and disposal obligations under the PRA, which govern what information is retained and how it is disposed of (destroyed or archived).
2. Stronger penalty regime. Other experts have called for increases to the maximum $10,000 penalty and where a penalty can be applied (Baker Wilson, 2026). This penalty can often be cheaper than security ("penetration") testing.
    - The Privacy Act binds the Crown, but very few acts of legislation specify how an Act is enforced against the Crown, reserved for high-stakes breaches like health and safety at work, or emergency management.

Given the responsiveness of the public service to existing independent officers, such as the Ombudsman, Privacy Act compliance actions should rarely escalate beyond remedial notices to an offence.
    - The 2026-2027 New Zealand Cyber Security Action Plan has directed the Ministry of Justice to investigate a stronger penalty regime to incentivise the protection of personal information (DPMC, 2026)
3. Requiring demonstration of privacy requirements, such as through the privacy management programme (PMP) concept outlined in the OECD privacy guidelines.
    - This recommendation is addressed in the public sector on a project-by-project basis through privacy impact assessments. No mechanism currently exists for demonstrating ongoing system-wide privacy protection to a similar extent as an organisational PMP. This concept is distinct from the system-wide security protections offered by the public-sector NZISM (as explained in Section 4.2.2). The NZISM does not distinguish between securing information collected legally, and securing information collected unnecessarily or disproportionately.
4. Providing stronger protections for automated decision-making, citing issues with "inaccurate predictions, discrimination, unexplainable decisions and a lack of accountability"
    - This recommendation is addressed in the public sector through both statutory provisions, explained in Section 4.1.5, and soft guidance, explained in Section 4.2.3. As explained in Section 4.2, the thesis recommends elevating this soft guidance to enforceable Public Service Act s57 guidance.

4.1.3.2. Privacy Act codes of practice

The OPC also develops codes of practice that prescribe guidance on the application of IPPs for data in specific industries and applications. These include civil defence and national emergencies, credit reporting, health information, and telecommunications, as well as exceptions regarding unique identifier sharing between justice entities and between superannuation schemes. Codes of practice are secondary legislation in which a breach of the code is treated as if the relevant IPP in the Privacy Act were breached. Codes of practice may strengthen or relax IPPs depending on the context. For example, the Health Information Privacy Code clarifies

when health privacy protections are stronger (e.g. parents have reduced access to information once a child is 16 or older) or weaker (e.g. legal representatives of deceased or incapacitated individuals may access information).

The most recently developed code of practice is the **Biometric Processing Privacy Code** (BPPC). The BPPC is the first code of practice that regulates the *processing* of personal information. Given that biometric processing systems are typically powered by machine learning (Tucci, Della Greca, Tortora, & Francese, 2024), the BPPC would be the first legislative measure to directly prohibit the use of any AI that fails to meet its criteria. The BPPC extends the threshold for necessary information collection under IPP 1 to test that biometric processing is: (bold emphases by OPC)

- **effective** at achieving such lawful purposes identified under IPP 1
- such purposes cannot be achieved by an alternative with less **privacy risk**
    - Privacy risk is defined as the likelihood of not only breaching the IPPs but also the risk of incorrect decisions based on an individual's personal attributes (i.e. biased decision-making), the risk of a chilling effect as a result of surveillance, the risk of information being reused for purposes unknown to the individual at the time, and that surveillance of individuals is avoided in "spaces where they may reasonably expect not to be monitored".
- **proportionate,** given the **benefit** of such lawful purposes **outweighs** the privacy risk
    - Where benefit is defined as the public benefit outweighing the privacy risk, the benefit to the data's subjects outweighing the privacy risk, or the private benefit substantially outweighing the privacy risk.
- Proportionality accounts for the cultural impacts and effects of biometric processing on Māori

The BPPC is unlikely to have a meaningful effect on the public sector, given that two major applications of biometrics, health and maintenance of the law, are exempt from the Code. Statutes in establishing legislation like the *Immigration Act 2009* s30, or the *Customs and Excise Act 2018* s203, already explicitly authorise the established agency's use of biometric information for specific uses. However, it may discourage the use of biometric collection in other parts of the public sector, such as in the education sector to monitor students and invigilate exams.

4.1.4. Automated electronic systems performing statutory actions

Legislation that establishes roles, responsibilities, and offences within a specific government domain may authorise an "automated electronic system" to carry out statutory actions outlined in such legislation. These systems are regarded by law as if the decisions were made by a comparable human in that statutory role. An automated electronic system is not necessarily an AI system; it often consists of straightforward business rules – some of which are delineated in the legislation itself – converted into code. However, these laws do not prevent the use of AI in the future if such AI meets the thresholds established by the legislation. A system may not fit this definition if it only informs a human who ultimately makes the decision.

The *Immigration Act 2009* marks the first reference in legislation to an "automated electronic system" with this distinction. Since 2012 (through the *Biosecurity Law Reform Act 2012*), a legislative pattern has emerged that includes the provision that the Chief Executive of the statutory officers may arrange for an automated electronic system if they are satisfied that "the system has the capacity to do the action with **reasonable reliability**" and that the automated decision can be **reviewed by a human** in the role being automated "without undue delay". The threshold for reasonable reliability has not been tested.

Offences against these statutory officers (such as intentionally obstructing or hindering) also apply to the automated system and introduce further offences for knowingly damaging or impairing it. The pattern allows the system to include components located outside New Zealand and requires consultation with the Privacy Commissioner when handling personal information.

Actions that can legally be automatically performed include:

- Immigration NZ applying predetermined criteria to process, grant or refuse an application for a visa or entry and to confirm New Zealand citizenship (*Immigration Act 2009*)
- Performing any power, function, or duty of a statutory role described in the *Biosecurity Act 1993*, *Wine Act 2003*, *Animal Products Act 1999*, and *Organic Products and Production Act 2023* (under the *Food Safety Law Reform Act 2018*)
- Customs performing their statutory powers to determine offences against regulation around the entry and exit of goods, persons and craft; forfeiture,

seizure and condemnation, among other miscellaneous powers (*Customs and Excise Act 2018*)

- Ministry of Justice's Legal Services Commissioner granting legal aid, with assessment criteria to follow, as prescribed in the *Legal Services Act 2011*
- Courts adding fines to existing arrangements (*Summary Proceedings Act 1957*)
- Altering MSD child support payments based on information shared by IR (*Social Security Act 2018*)

The *Social Security Act 2018* is the most evolved iteration of this pattern, mandating that a specific system for child support payments be under the department's control. It conditions the Chief Executive's satisfaction on consistency with an approved standard for the agency's (MSD) use of these systems, which are reviewed at least once every three years. This standard is explained further in Section 4.3.

This pattern does not appear in legislation where statutory powers are conferred on entities rather than on specified individuals. For example, ACC's empowering legislation does not include AES provisions, as 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 complex decisions that must be made by qualified assessors, such as determining rehabilitation support needs.

Courtney (2021), in his critique of the Algorithm Charter, suggests extending the right to a human alternative to automation and the right to review AES actions as general legislation covering all government automation. Courtney justifies this based on a lack of clarity regarding the applicability of AES authorising provisions, but it is unclear whether Courtney understood that the provision was part of a wider pattern. Given there is a clear pattern for automating specified actions, I argue that requiring legislative change for each AES use case compels agencies to rigorously debate the merits of each use case before justifying its inclusion in their Minister's legislative programme.

4.1.5. Appeals and reviews

A body of law known as 'administrative law' encompasses the previously mentioned provisions, such as the Ombudsman resolving complaints under both the

OIA and the Ombudsman Act, and the Waitangi Tribunal reviewing potential Treaty breaches, such as RoC*RoI.

Another aspect of the wider body of administrative law – judicial review – allows aggrieved parties to challenge the lawfulness of any decision (including AI-assisted ones) made by an executive agency through the court system. This process is established as a right under the *New Zealand Bill of Rights Act 1990* and is further established under the *Judicial Review Procedure Act 2016.* This right enables courts to assess whether a decision is lawful based on principles and precedents established by New Zealand courts. This process also considers relevant international precedents from similar legal systems where appropriate. Substantive matters of a decision itself are not examined in judicial review. Instead, the focus is on whether laws were adhered to in reaching a decision, such as whether an agency operated within its legislative powers (Crown Law Office, 2019). For example, judicial review may be conducted to determine the lawfulness of automating decisions in the absence of statutes that explicitly delegate this power, as described in Section 4.1.4 or in the context of other legal provisions, such as the right to non-discrimination under the *Human Rights Act 1993.* The Court of Appeal recently determined that government algorithms "must be amenable to [judicial] review" regardless of whether an algorithm exercises a statutory power of decision (McNicol, 2026). Any algorithm employed by the executive, regardless of whether it automates a specific statutory power, is open to judicial review.

4.2. Government-wide standards and guidance

Government agencies also adhere to best practice guidelines established by a "system lead" agency (for example, StatsNZ sets standards for data use as the data system lead). Figure 7 highlights the relevant guidance instruments in the guidance hierarchy. These guidelines carry varying degrees of legal authority. Most guidance is entirely voluntary, with some agencies publicly demonstrating their commitment to these guidelines as a form of accountability. Certain guidelines may become functionally binding for departmental chief executives through Cabinet or ministerial directives. Crown entities, which sit outside of the core public service, may also be bound to ministerial directives under section 107 of the *Crown Entities Act 2004.*

None of the standards or guidance mentioned in this section yet uses a system lead's powers under section 57 of the *Public Service Act 2020.* Section 57 enables system

leads to "set **standards** relating to the particular subject matter area that they lead and co-ordinate", which "apply only in or to public service agencies". System leads may also "issue **guidance** relating to that particular subject matter area", which "applies in or to all State services". Future system lead guidance should use these s57 powers to mandate compliance with algorithm and AI development practice.

This section introduces three categories of "standards": those endorsed by international standards like the ISO 27000 series, those that are nominally standards but derive a mandate from Cabinet directives (like the *Government Web Standards*) or secondary legislation (like the *information and records management standard*), and those that are nominally standards without an enforcement mechanism (like the *Digital Service Design Standard*).

This section also introduces the guidance instruments that take the form of **principles**: discrete strategic values that specify desired behaviours and outcomes at a high level, leaving agencies responsible for interpreting how the principles translate at a tactical level. As such, these principles have no legal mandate.

Guidance mentioned in this section includes instruments with actionable advice, such as those that are nominally 'tools' like the privacy, cloud or algorithm impact assessment. Others are framed as policies, such as the *Data Protection and Use Policy*; manuals, such as the *NZ Information Security Manual*; governance models devised by Te Kāhui Raraunga; and non-section 57 guidance, such as the *Responsible AI Guidance series*.

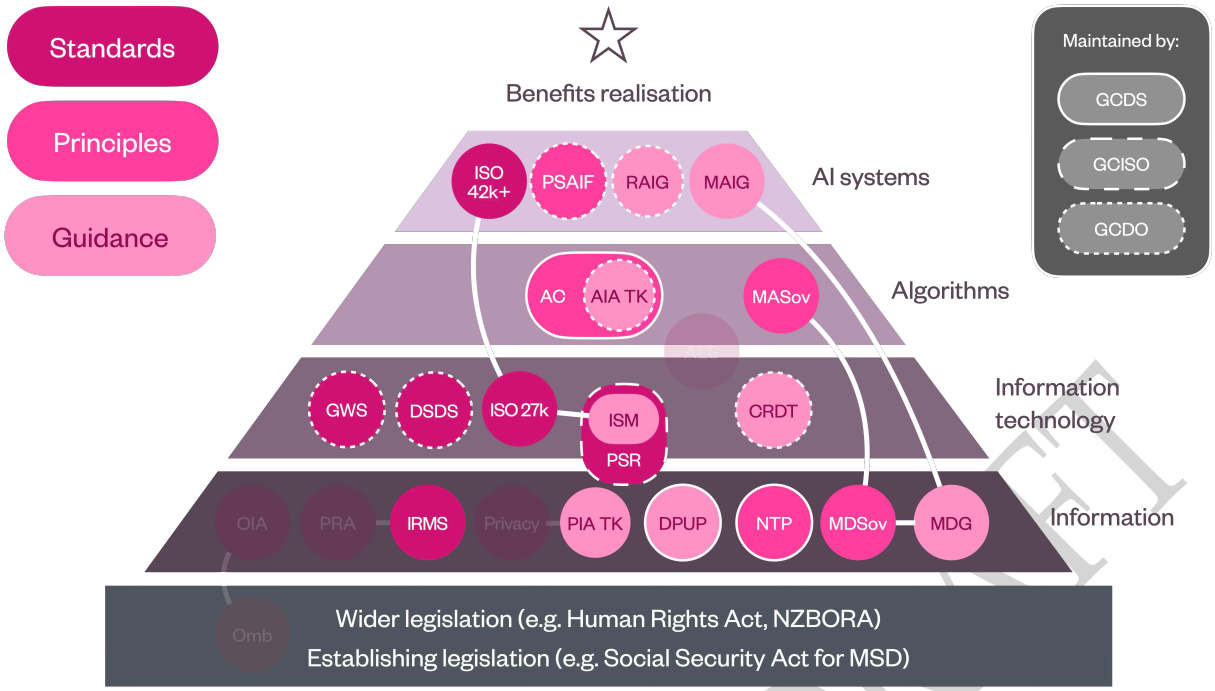

Figure 7: Standards, principles and guidance relevant to AI system delivery highlighted in the guidance ecosystem hierarchy.

4.2.1. Information guidance

Figure 8 highlights relevant guidance instruments at the information level – governing the use of people's personal data. Two information guidance instruments have already been mentioned in this chapter, given their close association with specific legislation: the Privacy Commissioner's privacy impact assessment toolkit and Archives New Zealand's information and record management standard.

StatsNZ, in its role as the Government Chief Data Steward (GCDS), develops government-wide guidance for the use of data by government. Two of the instruments mentioned in this section are maintained by StatsNZ: the Data Protection and Use Policy and Ngā Tikanga Paihere. The remaining instruments focus on the use of Māori data: with Māori data sovereignty principles devised by Te Mana Raraunga, an academic network, and the Māori Data Governance Model devised by Te Kāhui Raraunga, a working group governed by iwi leaders.

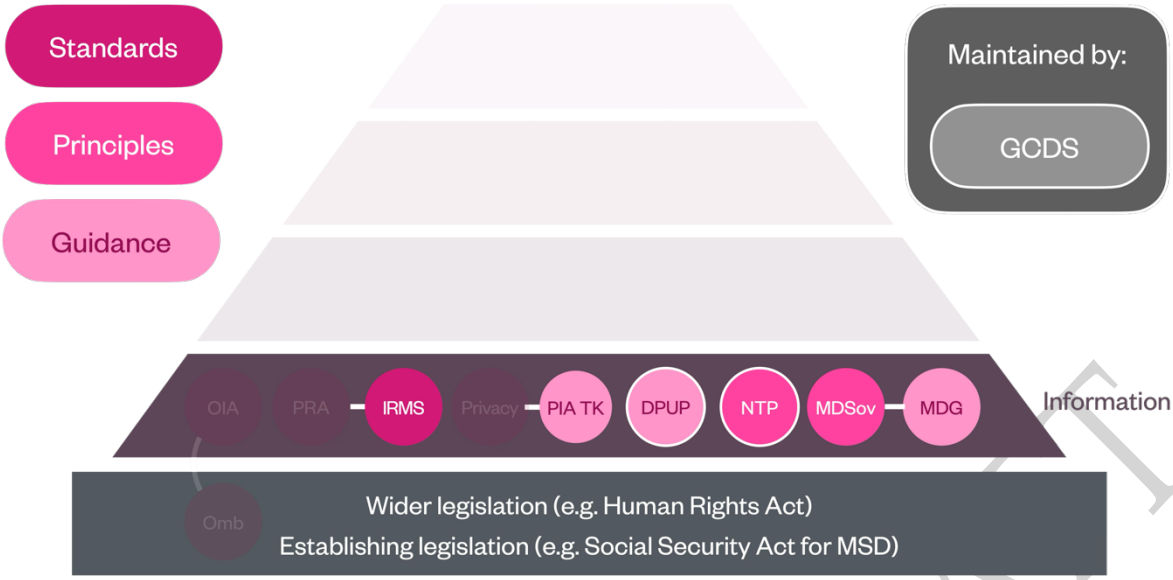

Figure 8: Guidance instruments relevant to any collection and use of information.

4.2.1.1. Principles for safe and effective use of data and analytics (PSEUDA)

The first framework specifically designed to guide data use in government was the "**Principles for safe and effective use of data and analytics**" (StatsNZ and OPC, 2018). These principles – outlined below – establish a foundation for how government agencies use data and analytics responsibly.

- Deliver clear public benefit: data use considers and demonstrates the benefit of using individual data.
- Maintain transparency: foster collaboration and share responsibility.
- Understand the limitations: all analytical processes have inherent constraints.
- Retain human oversight: analytics should not replace human decision-making
- Ensure data is fit for purpose: use the appropriate data in the correct context to prevent adverse outcomes.
- Focus on people: remember the individuals behind the data who expect their data to be protected against misuse.

These principles form the basis for the subsequent Algorithm Charter (explained in Section 4.2.3). The Charter acknowledges that it works alongside the PSEUDA and does not replace it. Despite this, no recent all-of-Government directives refer to these principles.

4.2.1.2. Integrated data safety

As StatsNZ maintains the Integrated Data Infrastructure (IDI) used by government agencies to systematically integrate various sources of government data, its internal **Five Safes framework** governing the safety of the IDI must be adhered to by all its users. (StatsNZ, 2022). This framework primarily addresses privacy and security considerations concerning the safety of a microdata research project (safe people, safe projects, safe settings, safe data, safe output).

StatsNZ maintains a distinct framework for governing the use of Māori data in the IDI known as **Ngā Tikanga Paihere** (NTP), drawing on tikanga (the customary system of values and practices in the Māori world) to inform the culturally appropriate use of Māori data within the IDI (StatsNZ, 2020). This framework requires agencies to:

1. Have appropriate expertise, skills, and relationships with communities
    - Pūkenga: Researchers demonstrate awareness of, and an intention to work with, data in culturally appropriate ways.
    - Whakapapa: Researchers establish suitable relationships with communities before undertaking substantive research.
2. Maintain public confidence and trust in using data
    - Pono: Researchers demonstrate an awareness of and intention to work with data in culturally appropriate ways.
    - Tika: The level of accountability to communities of interest is explained, and there is community support for the research.
3. Use good data standards and practices
    - Wānanga: Institutions have established systems and procedures to support culturally appropriate data practices.
    - Kaitiaki: Communities of interest are identified and involved in research decisions as early as possible.
4. Have a clear purpose and action
    - Wairua: Community objectives align with research objectives, and any potential harm is considered.
    - Mauri: Researchers show how data transforms from its original collection purpose to support research objectives.

5. Balance benefits and risks
    - Tapu: Sensitivities in the use of data are identified, including privacy issues for whānau and identifiable groups.
    - Noa: Data is readily accessible, and there is demonstrated awareness of the impact on communities of interest.

The Government Data Strategy (StatsNZ, 2021) has identified NTP's implementation "across the data system" as an important outcome in fulfilling the Crown's responsibilities to Te Tiriti o Waitangi. StatsNZ has not publicly reported on the progress or success of this implementation.

4.2.1.3. Māori data sovereignty

While NTP outlines some obligations that government agencies have under Te Tiriti o Waitangi, it was drafted as a cultural appropriateness framework. NTP does not focus on giving effect to Te Tiriti o Waitangi – particularly the obligation to allow Māori to exercise sovereignty over their taonga. Instead, Te Mana Raraunga (TMR), an academic collective of experts in Māori data and research, has developed the Māori Data Sovereignty (MDS) Principles (Te Mana Raraunga, 2018) to uphold the rights and interests to which Māori people, resources, environments, and taonga are entitled under Te Tiriti o Waitangi. The principles include:

1. Rangatiratanga (Authority)
    - Right to control Māori data and data ecosystems
    - Store data in Aotearoa New Zealand to enhance control
    - Right to access relevant data to empower Māori to self-govern
2. Whakapapa (Relationships)
    - Metadata concerning the provenance and context of the data should be accessible to understand its whakapapa (genealogy, origins)
    - Data categorisation and measurement should prioritise Māori aspirations
    - Prevent future harm from the use of current data
3. Whanaungatanga (Obligations)
    - Individual rights are balanced with those of the collective, which prevails in some contexts
    - All actors involved in every aspect of the lifecycle of Māori data use are accountable to the individuals whom the data represents

4. Kotahitanga (Collective benefit)
    - Empower Māori to derive individual and collective benefits
    - Develop a Māori workforce proficient in all aspects of the data lifecycle
    - Connect with other indigenous communities to share valuable insights
5. Manaakitanga (Reciprocity)
    - Respect and uphold the dignity of Māori from whom the data is sourced; insights that stigmatise or blame Māori should be avoided
    - Require free, prior and informed consent
6. Kaitiakitanga (Guardianship)
    - Empower Māori to exercise kaitiakitanga in data storage and transfer
    - Māori protocols and knowledge underpin the data lifecycle
    - Māori determine the restriction or publication of Māori data.

Within this framework, NTP is only directly applicable under the sixth principle, where Māori protocols (including tikanga) must underpin all aspects of the data lifecycle. However, these protocols are based on the same tikanga concepts. The only aspect of MDS not present in NTP relates to rangatiratanga, where StatsNZ, as an agent of the Crown, does not explicitly cede control over Māori data to Māori.

TMR has also developed the Māori Data Audit Tool, which guides organisations step-by-step in assessing their ability to address the Māori data sovereignty principles. No government agency publicly discloses how or if it implements TMR's model in its standard operating procedures. However, StatsNZ has agreed with a different body – the Data Iwi Leaders Group (DILG) – to commit to policies that align with the aspirations of MDS. As its name suggests, DILG is an iwi-led initiative rather than an academic one like TMR.

4.2.1.4. Māori Data Governance model

Te Kāhui Raraunga, the operational arm of the DILG convening technical experts to give effect to DILG's aspirations, has designed an operating model for Māori Data Governance (MDG), the operational standards and policies that give effect to MDS (Te Kāhui Raraunga, 2023). Table 4 visualises this model's overarching vision, desired outcomes, and guiding values; an authority mechanism that translates these aspirations and authorises actions under seven Data Pou (pillars), with a foundational eighth Data Pou enabling the implementation of these seven consistently. This model is the basis for the Māori AI Governance Model (Te Kāhui Raraunga, 2025).

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).

The MDG model directly references the use of algorithms and AI as a significant category of data use (as guided by Data Pou 6). In addition to MDS considerations, the MDG model also specifically recommends "at a minimum that Māori should have the right to…

- know whether their data is being used to develop and/or train machines or algorithms and
- be free from data practices that are deceptive, manipulative, coercive, discriminatory and that cause harm to individuals or groups, whether that harm is intended or not, and
- interrogate and influence data practices and processes that affect them, including operational algorithms." (Te Kāhui Raraunga, 2023)

4.2.1.5. Data Protection and Use Policy

StatsNZ's Centre for Data Ethics and Innovation has adopted the Data Protection and Use Policy (DPUP). It was originally developed by the Social Wellbeing Agency, then transferred to the Government Chief Privacy Officer (GCPO) following its broader endorsement across government, until the GCPO was disestablished. DPUP was devised to enable the creation of authentic relationships and trust with the different communities the Government and contracted non-governmental organisations serve.

DPUP provides normative commitments on "doing the right thing […] when collecting or using people's data and information" through five principles that outline values needed to act as responsible users of personal data and four guidelines that describe how to give effect to those principles, as summarised in Table 5.

Table 5: DPUP principles and guidelines (Social Wellbeing Agency, 2021)

Principles	Guidelines
He Tāngata – Focus on improving people's lives: individuals, children and young people, whānau, iwi and communities. **Manaakitanga** – Respect and uphold the mana and dignity of the individuals, whānau, communities or groups who share their data and information. **Mana Whakahaere** – Empower people by giving them choice and enabling their access to and use of their data and information. **Kaitiakitanga** – Act as a steward in a way that people understand and trust. **Mahitahitanga** – Work as equals to create and share valuable knowledge.	**Purpose Matters** – be clear about the purpose for collection (only what is needed), use, and sharing. **Transparency and Choice** – help people understand why and how providing personal information can help them or others in similar circumstances, what is optional for them, and their rights to access or request changes. **Access to Information** – support people in understanding what information is held about them **Sharing Value** – include and involve people with the right experience, and confirm and share insights with those who could benefit

Thus, DPUP goes further than both existing legal obligations and frameworks such as the Algorithm Charter. For example, the He Tāngata principle provides the clearest example of what delivering clear public benefit looks like, as per the PSEUDA: by improving people's lives.

4.2.2. Information technology standards

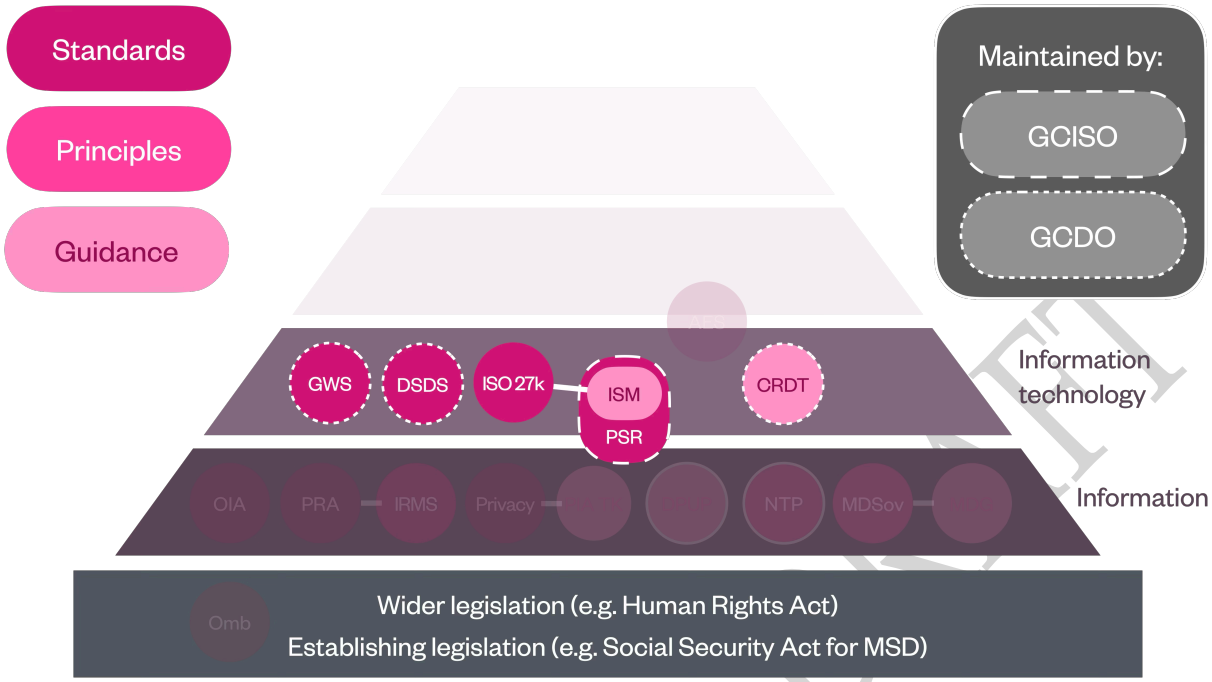

Figure 9: Guidance instruments related to the development of any information technology.

Figure 9 highlights the guidance instruments developed by two system leads to guide the general use of information technology. Currently, DIA leads the coordination of an all-of-government approach to the purchasing and use of digital technologies, including AI, as discussed in Section 4.2.4. As such, the chief executive of DIA also has the title of Government Chief Digital Officer (GCDO). The GCDO coordinates government procurement of common digital products and services, allowing agencies to easily purchase commonly used tools. For example, the Government has one contract with DXC Technology for facial recognition as a service, including the uses explained in Section 3.1.3. Agencies can join this existing contract without conducting as extensive due diligence as they would going out to market independently (NZGP, 2019).

The GCDO also maintains standards regarding usability and accessibility in designing digital products and services. For example, their **Government Web Standards** mandate that websites (such as chatbots and assistive web search) must be navigable by keyboard and assistive technologies (like screen readers and magnifiers) and use sufficiently contrasting colours. The GCDO's non-mandatory **Digital Service Design Standard** offers 12 principles for best practice in digital service design, including principles that overlap with algorithmic guidance. These standards are

mandated by the Government Procurement Rules for any government agency procurement of digital technology.

A different agency leads all-of-government practice for cybersecurity. The chief executive of the Government Communications Security Bureau also acts as the Government Chief Information Security Officer (GCISO). The GCISO develops and maintains **the New Zealand Information Security Manual** (NZISM). This manual gives effect to the legally mandated Protective Security Requirements that govern all aspects of government security. It establishes necessary protections for the use of information and information technology within the government. The manual draws on the **ISO 27000 family of standards**, which standardises and provides guidance on managing risk in information systems and is typically already implemented globally by both government and non-government entities.

The manual provides guidance from the highest level, including the processes for certifying and accrediting systems and responding to cybersecurity incidents, down to lower-level considerations that regulate the use of specific technologies such as cordless telephones, cryptographic algorithms, and cloud computing services. Currently, there is no specific guidance on the use of AI in the NZISM.

The GCSB also hosts the National Cybersecurity Centre (NCSC), whose primary function – as the nation's computer emergency response team – is to assist organisations, both public and private, in cybersecurity incident management. The NCSC also works with its allied counterparts to issue non-binding joint guidance on the secure use of AI for any NZ organisation to consider.

4.2.3. Algorithm guidance

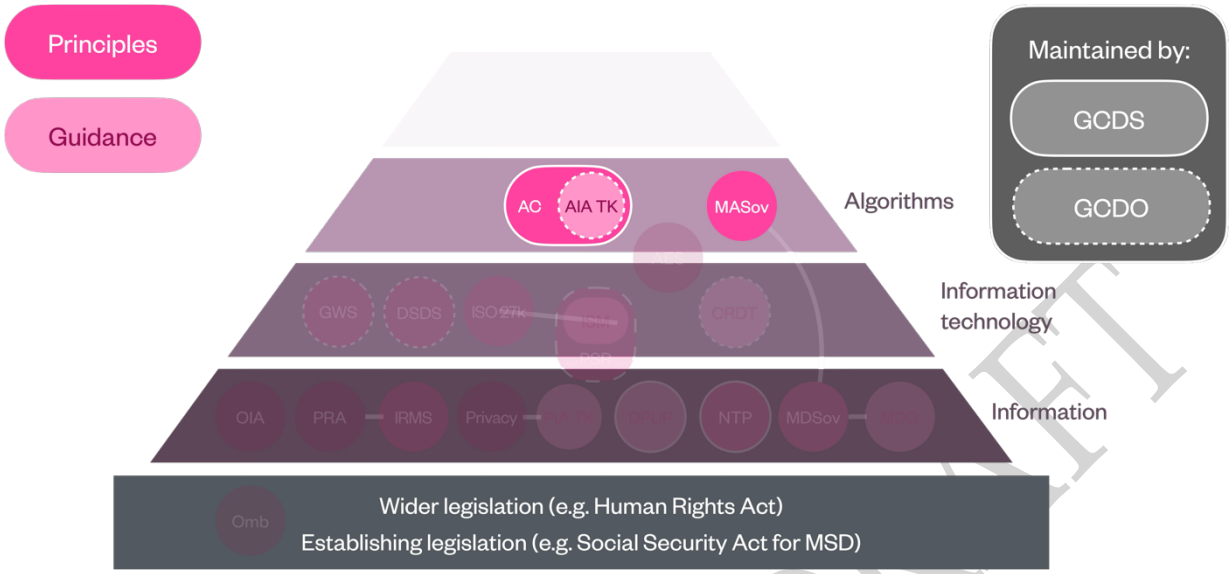

Figure 10: Guidance instruments related to the development of algorithmic systems

There are only two bodies of guidance at the algorithm level, as highlighted in Figure 10. The only framework known to presently govern the Government's use of algorithms and AI is the Algorithm Charter for Aotearoa New Zealand ("the Charter"), a voluntary commitment by signatory agencies to ensure that the algorithms they employ are suitable for their intended purpose and comply with legal and ethical standards (StatsNZ, 2020). These commitments mostly align with the 2018 data and analytics principles, with the Charter being less detailed but more understandable to the public. Some similar principles have key differences:

- Maintaining transparency also involves making information about the data and processes available, notwithstanding existing legal restrictions.
- Delivering a clear public benefit is only seen as an outcome of fulfilling Treaty commitments and integrating a Te Ao Māori approach.
- Understanding the limitations of the data is further elaborated in a new commitment to safeguard privacy, ethics, and human rights through regular peer reviews.

As of March 2026, most government agencies that employ algorithms have signed the Charter, along with several notable exceptions:

- **Te Whatu Ora Health New Zealand**, formed by the 2022 incorporation of all district health boards into one organisation, is now responsible for all public

health service delivery. Algorithms play an important role in health service delivery, as discussed in Section 3.1.2.

- **Kāinga Ora**, the government's state housing provider, which was reorganised into its existing form a year before the Charter was signed. However, the actual decisions around state housing eligibility are handled by a different agency: MSD, a signatory of the Charter. Little is publicly known about any algorithms or AI they may use for internal purposes.
- **Treasury and Reserve Bank**, both of which are exclusively policy agencies that adjust fiscal (taxes and government spending) and monetary (ease of lending) policy, respectively. They both extensively employ algorithms to inform their decision-making, as discussed previously. The Reserve Bank also sits outside the definition of the public service used in system lead guidance.
- **Customs Service**, whose enabling legislation explicitly allows it to use "automated electronic systems" to carry out any of its statutory roles, as discussed in Section 4.1.4, in its role of protecting the nation against threats that may come into or leave its border.
- **Ministry of Defence**, which advises the Government on defence policy. The New Zealand Defence Force, which comprises the three operational branches of the NZ military, is a signatory and is actively implementing the Charter and the AIA toolkit in its governance processes. (Woods, 2024).
- The **intelligence sector** has its own line of general accountability for its operations through the Inspector-General of Intelligence and Security. Due to the highly classified nature of their operations, they are typically not involved in open government initiatives such as the Charter.

4.2.3.1. Algorithm impact assessment

The Algorithm Impact Assessment (AIA) toolkit was released in late 2023 to serve a similar purpose as the PIA toolkit in guiding privacy compliance for the Charter (Tweedie, 2023). The AIA toolkit provides a definition that officially supersedes the scope of the Charter, simplifying it by requiring that automated decisions having a material impact on individuals or groups use the AIA process. This definition also clarifies the original design of the Charter, encompassing all AI and ML. Consequently, the definition anachronistically includes GenAI. The AIA process offers much more detailed guidance on end-to-end algorithm development, providing

prompts to support thinking from the business case through monitoring and evaluation.

The AIA was likely inspired by the closest international cognate to the Algorithm Charter: the Canadian federal government's *Directive on Automated Decision-Making (DADM)* (Chen A. , 2022). The DADM, however, is compulsory for any ADM system employed by the Canadian federal government. Furthermore, unlike New Zealand's AIA, Canada's AIA is quantitative: higher questionnaire scores indicate higher impact, with mitigations lowering the score. Canada's AIA is much more comprehensive than New Zealand's, interrogating an automated decision's effects on the economic interests of individuals or communities, on the sustainability of an ecosystem, and on the degree to which such a decision is reversible and lasting (Treasury Board of Canada Secretariat, 2023).

Unlike New Zealand's algorithmic guidance, the D-ADM prescribes actions based on four tiers of increasing risk scores. For example, an impact level I system will not require expert peer review, impact level II and III systems will require at least one expert to peer review, and impact level IV systems require two separate experts. New Zealand's Algorithm Charter only prescribes actions across all algorithms.

4.2.3.2. Māori algorithmic sovereignty

Brown et al. (2024) argue that since modern algorithms – particularly those that make decisions for Māori people and taonga – can be constructed by analysing Māori data, the principles that support Māori data sovereignty should be extended to enable Māori sovereignty in the algorithm development lifecycle, collectively referred to as **Māori algorithmic sovereignty (MAS)**:

1. Rangatiratanga: Māori have the right to oversee the development and use of an algorithm in a way that enhances their self-determination, including motives, management, and storage jurisdiction.
2. Whakapapa: Māori are aware of all aspects of the data throughout the algorithmic system, and its application provides broader benefits to the environment in which the data originates.
3. Whanaungatanga: the perspective of the collective needs to be considered alongside that of the individual, recognising the Tiriti principle of the right to redress and the ability to challenge an algorithm's outcome.

4. Kotahitanga: algorithms enable Māori to attain both individual and collective advantages while minimising harm.
5. Manaakitanga: the use of algorithms respects and upholds the mana and dignity of Māori, including their privacy and informed consent.
6. Kaitiakitanga: Māori are empowered to serve as kaitiaki (loosely translated as protectors) of all aspects of the algorithm, with tikanga (including categorising inputs and outputs as tapu/restricted or noa/accessible), kawa, and mātauranga underpinning this protection.

Brown et al.'s novel contribution to this guidance ecosystem lies in conceptualising an algorithm as part of a broader system. They argue that an algorithmic system is guided and constrained by the algorithm's motives and design to fulfil a specific process, as illustrated in Figure 11. While they advocate for assessing the alignment of each element of existing algorithms with the six MAS principles, they prefer that new algorithms be developed with motives and design firmly and unambiguously rooted in the tikanga values presented by the MAS principles.

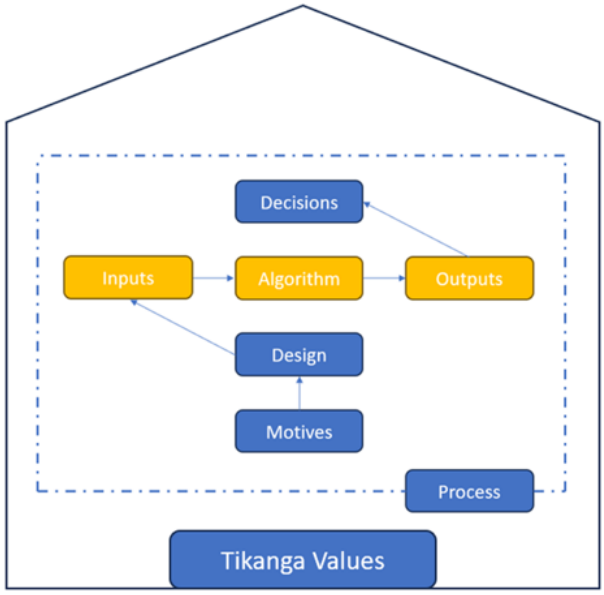

Figure 11: In yellow – the current concept of an algorithm, which may be guided or limited by the influencing elements in blue. An "indigenised" algorithm is situated within a structure based on tikanga values. (Brown et al., 2023)

4.2.4. AI guidance

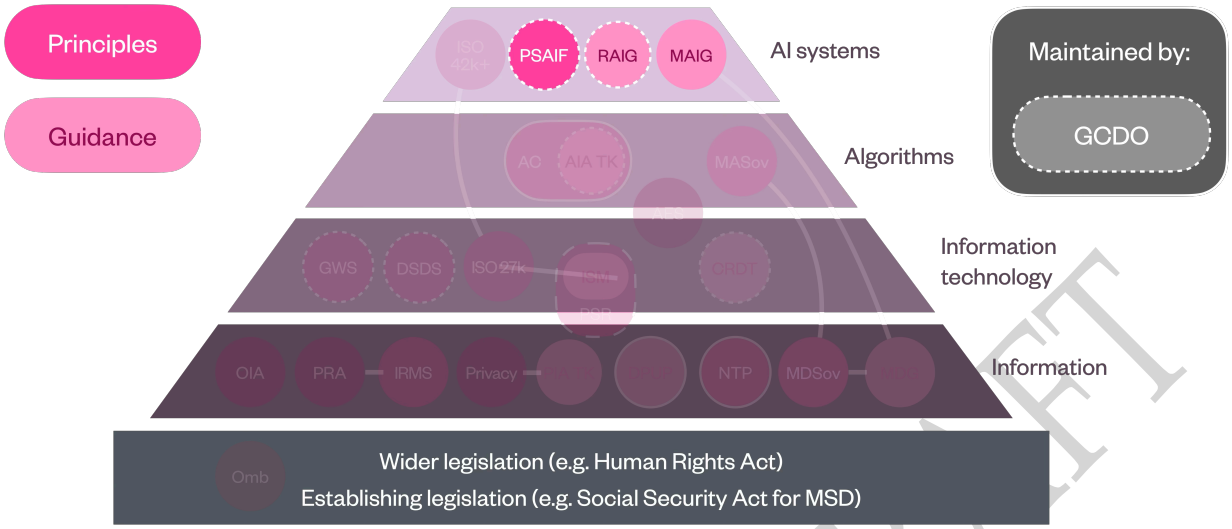

Figure 12: Guidance instruments related to the development and use of AI systems

Figure 12 shows that most guidance instruments at this level of the hierarchy are developed by the GCDO. Unlike the Charter, which addressed algorithms that use (if not derive from, as in machine learning) personal data, AI guidance arises from the GCDO's mandate to coordinate the procurement and use of technology platforms. The GCDO initially issued interim guidance around how agencies govern GenAI tools and platforms, along with broad recommendations for how end users should interact with AI. This guidance has since evolved into the Responsible AI Guidelines (RAIG) for the Public Service: GenAI, which links more fulsome guidance to the OECD AI Principles. The RAIG is intended to be a series, with the Ministry of Business, Innovation and Employment releasing a further RAIG for businesses as part of their mandate in microeconomic policy.

In parallel with this guidance, the GCDO has also developed the **Public Service Artificial Intelligence Framework** (PSAIF), a strategic framework that outlines a vision for how the New Zealand public service should adopt AI. It adopts and modifies the *OECD AI Principles* and links them to the unique New Zealand context, which "already provides some guardrails" (DIA, 2025). Table 6 demonstrates the differences between the PSAIF and OECD AI principles.

Table 6: The five OECD AI principles for the trustworthy use of AI, along with a summarised explanation of the application of each principle and its relationship with the Charter principles (OECD, 2019)

PSAIF principle	OECD AI principle	Gaps with existing laws and guidance
Inclusive, sustainable development	Inclusive growth, sustainable development and well-being	The equivalent PSAIF principle **falls short** of the OECD principle, removing "wellbeing". DPUP still promotes wellbeing as an artefact of the Ardern government. No guidance instrument mentions a focus on beneficial outcomes for planet in addition to people.
Human-centred values	Human rights and democratic values, including fairness and privacy	The equivalent PSAIF principle **falls short** of the OECD principle, removing "freedom, […] diversity, fairness, social justice, and internationally recognised labour rights".
Transparency and explainability	Transparency and explainability	This principle is covered in the PSAIF and the Charter and is supported through the Official Information Act.
Security and safety	Robustness, security, and safety	This principle is covered in the PSAIF and supported by security standards and guidance.
Accountability	Accountability	The equivalent PSAIF principle goes further than the OECD principle, focuses on human oversight "with appropriate […] capability" as the main accountability mechanism, presumably with the capability to withstand automation bias.

While the PSAIF was launched alongside the *RAIG for the Public Service: GenAI* (RAIG-PSG), they read as independent documents rather than as deliberately integrated and interdependent. For example, the RAIG-PSG has opted to use the *OECD AI Principles*, rather than the principles of the PSAIF, which have been "inform[ed]" by, but are simpler than, the OECD principles. 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.

The GCDO has committed to a two-year plan of AI initiatives called the Public Service AI Work Programme (DIA, 2026). In this work plan, they commit to developing a public service AI assurance model and toolkit, as well as standardised public service AI certification mechanisms.

4.2.5. Mapping guidance to legal obligations

It is important to understand whether individual aspects of 'soft' guidance offered by the instruments enumerated in this section currently have a legislative or broader legal mandate. Table 7 first groups aspects of existing guidance by common themes in the left-hand column. This grouping demonstrates significant overlap across guidance instruments. From there, in the right-hand column, I comment on whether this particular aspect of guidance currently has a legal mandate and whether a lack of mandate is desirable to enable sufficient adaptability.

Table 7: Aspects of practice offered in existing guidance and whether such suggested practice has a legal mandate, is not currently (or only partially) mandated but should be fully mandated by standards, or is only partially mandated which is desirable.

Aspects of existing guidance	**Comment on legal mandate**
<u>Strategic priorities</u>	
PSEUDA: "Deliver clear public benefit." vs **DPUP**: "He Tangata: Focus on improving people's lives" vs **PSAIF:** "Inclusive sustainable development: focus on innovation, efficiency and resilience."	Not mandated, **requires mandatory standards only**: Agencies only have obligations insofar as the *Public Service Act 2020* sets out how the public service should conduct its affairs, held accountable by their Minister or board, but this Act's principles do not specify overarching strategic policy goals to the same extent as these guidance instruments. It is desirable for strategic policy goals to be articulated in public service guidance within an objective normative framework devised by an impartial public service, while recognising the need to serve the priorities of the government of the day.
<u>Bias identification and management</u>	
PSEUDA: "Ensure data is fit for purpose" and **AC**: "Make sure data is fit for purpose by identifying and managing bias."	Mandated: Agencies have obligations under the *Human Rights Act 1993* and risk being investigated under the *Ombudsman Act 1975* if actions are being performed by systems that discriminate based on the protected attributes defined in the Human Rights Act, or on irrelevant grounds as determined by the Ombudsman.
<u>Community engagement</u>	

Aspects of existing guidance	Comment on legal mandate
PSEUDA: "Focus on people", **AC**: "Focus on people by identifying and actively engaging with people, communities and groups who have an interest in algorithms, and consulting with those impacted by their use."	Not explicitly mandated, **requires mandatory standards only**: Agencies only have obligations insofar as the *Public Service Act 2020* sets out that the public service must act "with a spirit of service to the community". DPMC (2025) notes that some agencies require community engagement. This practice is a matter regarding wider organisational culture, rather than an issue specific to algorithms and AI, and may be addressed through standards of conduct issued by the PSC.
Respecting communities	
DPUP: "Manaakitanga: Respect and uphold the mana and dignity of the people, whānau, communities or groups who share their data and information."	Not explicitly mandated, **requires mandating DPUP**: As above, as well as the public service values including "respectful: to treat all people with dignity and compassion and act with humility" as enforced by PSC codes of conduct.
Ease of access	
DPUP: "Mana Whakahaere: Empower people by giving them choice and enabling their access to and use of their data and information."	Mandated by the *Privacy Act 2020* and *Official Information Act 1982*, but DPUP notes "people should not have to rely on […] requests to access information held about them" and consider proactive ways to provide the mandated right to access and correct.
Community collaboration	
DPUP: "Mahitahitanga: Work as equals to create and share valuable knowledge."	Not mandated, **requires mandating DPUP**: As for the "focus on people" principle, this practice is a matter regarding wider organisational culture, rather than an issue specific to algorithms and AI, and may be addressed through standards of conduct issued by the PSC.
Transparency	

Aspects of existing guidance	Comment on legal mandate
PSEUDA: "Maintain transparency" and **AC**: "Maintain transparency by clearly explaining how decisions are informed by algorithms" and **PSAIF**: "Transparency and explainability"	Partially mandated, desirable: Agencies have obligations under the *Official Information Act 1982* s22-23 where systems are making personally impactful decisions (consistent with the algorithm threshold assessment that defines when the AC is used), and under s12 for the general use of algorithm and AI systems. The scope of s22-23 acts as a risk threshold, mandating expanded transparency for high-risk systems, and baseline transparency for all other systems.
Suitability of data	
PSEUDA: "Understand the limitations" and **AC**: "Make sure data is fit for purpose by understanding its limitations"	Not mandated, **requires mandatory standards only**: The limitations of data are measured through technical methodology – there is no one threshold that deems data to be accurate and representative enough to be reliably used in a system. This issue is best addressed through mandated technical standards.
Privacy, human rights, and ethics	
AC: "Ensure that privacy, ethics and human rights are safeguarded" and **PSAIF**: "Human-centred values"	Mandated: Agencies have obligations under the *Privacy Act 2020*, and *Human Rights Act 1993*. Ethics is too broad to be legislated and is better articulated through guidance.
Human oversight and accountability	
PSEUDA and **AC**: "Retain human oversight" and **PSAIF**: "Accountability"	Mandated: Agencies must legislate where statutory actions performed by specified persons are being automated: to provide for a human alternative without undue delay. Appeals and reviews are always available, if not through a specific channel under their enabling legislation, through the *Ombudsman Act 1975*.
Integrated data safety	

Aspects of existing guidance	Comment on legal mandate
Five Safes framework	Mandated for any StatsNZ research data access under the *Data and Statistics Act 2022* s48. Not mandated for any other use of data.
<u>Culturally appropriate data collection and use</u>	
Ngā Tikanga Paihere	Not mandated, other more fundamental principles as below should be mandated instead. NTP is only a framework for ways of working with tikanga applied as appropriate where needed, case by case. More fundamental guidance in the two rows below should be mandated instead.
<u>Māori sovereignty rights</u>	
Māori data sovereignty, Māori algorithmic sovereignty (MDS/MAS)	Constitutionally mandated but no enforcement mechanism, as Waitangi Tribunal findings are not legally binding. **Māori sovereignty over specific taonga can and should be legislated** for within existing constitutional arrangements.
<u>Operating models for collecting and using Māori data</u>	
Māori Data Governance Model, Māori AI Governance Model	Not mandated, **requires mandatory standards only**. The aims of Māori data sovereignty principles may be achieved by mandating this governance model, without explicitly legislating for Māori data sovereignty.
<u>Digital accessibility and usability</u>	
Government Web Standards	Mandated by Cabinet minute (03) 41/2B
<u>Service design</u>	
Digital Service Design Standard	Not mandated, may be mandated as it matures. This standard acts more as a maturity framework for service design practice within agencies, rather than critical guidance for IT development. It is currently being

Aspects of existing guidance	Comment on legal mandate
	redeveloped; thus, a more mature standard may be legally mandated as a s57 standard.
Protective and information security	
Protective Security Requirements and NZ Information Security Manual, PSAIF: "Safety and security", **DPUP**: "Kaitiakitanga: Keep data and information safe and secure and respect its value"	Mandated by Cabinet minute (14) 39/38

This table shows that only one critical aspect of guidance is not specifically legislated for: Māori data sovereignty. While potential breaches of Te Tiriti o Waitangi related to the use of Māori data can be brought before the Waitangi Tribunal, the Tribunal's findings are not legally binding. Regardless of whether there is an enforceable mandate, agencies still have constitutional obligations pursuant to the Waitangi Tribunal's recognition of Māori data as having potential to be taonga, which must be actively protected by the Crown (Waitangi Tribunal, 2021). Thus, it would be advisable for mandatory standards to enable agencies to understand such Treaty obligations.

All other aspects of guidance are either legally mandated, not/partially mandated but should be elevated to s57 standards rather than legislation to remain flexible to technological or political developments, or only partially mandated but desirable.

4.3. Internal agency policies

The "soft" government-wide guidelines on data, algorithms, and AI use mentioned in Section 4.2 gain more "local" enforcement power when integrated into the enterprise policies that govern how an agency approaches specific processes. Deviation from the guidelines now carries the risk of internal consequences but depends on internal enforcement unless a provision for external review exists.

4.3.1. Ministry of Social Development

One agency with extensive proactive transparency around algorithm policies is MSD. As one of the larger government agencies, MSD is responsible for a wide range of administrative decisions, from child support payments to student loans to superannuation. Therefore, MSD benefits greatly from automating these decisions and from higher-level policymaking that determines the provision of interventions.

As previously mentioned, the Social Security Act (SSA), which empowers MSD's statutory functions, provides legal enforcement for an internally developed Automated Decision-Making Standard (ADMS). The ADMS encompasses MSD's Algorithm Charter commitments and reiterates its legal obligations under the SSA and OIA, while also extending them in several ways (Vowles, 2021):

- Biases identified as part of the Charter commitment must be removed or acceptably mitigated. The Charter only requires that biases be managed.
- The differential risk of fraud in a new ADM must be assessed, and any potential increase in that risk must be mitigated.
- Reviewing an automated decision cannot be automated; it must be performed by a human.

MSD has two additional frameworks that must be considered during model development. Their Model Development Lifecycle (MDL) is available on their website, which goes further than the AIA toolkit by providing both technical guidance and enterprise governance guidelines at each stage of the model development process (Ministry of Social Development, 2021). The legally binding ADMS mandates the use of the MDL for complex algorithms. The ADMS and MDL do not refer to GenAI techniques, as they predate the advent of mass-market GenAI.

MSD also enforces the Privacy, Human Rights, and Ethics (PHRaE) framework when collecting, using, or disclosing personal information. PHRaE aligns data use with legislation, including the *Privacy Act*, the *Human Rights Act*, and *Te Tiriti o Waitangi* (Sepuloni, 2018). The complete PHRaE framework is no longer available to the public.

4.3.2. Health New Zealand Te Whatu Ora

Health New Zealand Te Whatu Ora (HNZ) recognises that AI and algorithm safety in the health sector "went beyond simply adapting existing risk frameworks and governance guidance to [NZ's] specific health service context and population" (Whittaker et al., 2023). Consequently, they have developed a governance framework that includes other pertinent aspects of AI in health service delivery, as illustrated in Figure 13.

1. The first consideration is the overarching appropriateness in the proposed context, which involves determining whether "the problem at hand is clearly defined and well understood [...] and the data required for AI development [...] is of sufficient scale and quality".
2. Consumer perspectives on their trust in the purpose, beneficence, and safety of AI applications are enhanced by transparency and communication.
3. Māori perspectives on what systems will look like, ensuring that these systems perform effectively for Māori and safeguarding their sovereignty as per Te Mana Raraunga (2018).
4. Equity and fairness require extending performance considerations to other groups at risk of biased decision-making.
5. Ethical considerations, while relatively underdeveloped in existing general AI governance frameworks, are already well established in the NZ health system through frameworks such as the National Ethical Standards. They prioritise autonomy (allowing patients to decide who – or what – makes their decisions) as well as beneficence and non-maleficence (ensuring that benefits outweigh risks).
6. Operational perspectives of clinicians who use AI primarily centre on human oversight and explainability, as in other general AI governance frameworks.
7. Data issues surrounding data completeness and its representativeness of the target population.
8. Technical guidance as per existing standards (like those identified in Section 4.2.2).
9. Legal and privacy perspectives across laws (like those identified in Section 4.1) and the Code of Health and Disability Services Consumers' Rights.

Figure 13: Illustration of Health New Zealand Te Whatu Ora's AI assessment framework (Jin, 2024)

HNZ's AI governance framework provides two novel contributions. Its integration of domain-specific considerations into its overarching governance policy demonstrates where all-of-Government guidance will fall short due to its general nature. Specialist agencies will need to tailor general guidance based on their unique operating environment and risk profile. Health NZ's framework is also underpinned by interrogating appropriateness, not just the fundamental appropriateness of AI as a solution to a defined problem, but appropriateness across all domains. Notably, neither the GCDO's PSAIF nor StatsNZ's Algorithm Charter treats appropriateness as an explicit governance principle, despite this concept being a core determinant of trustworthiness.

4.3.3. Accident Compensation Corporation

The Accident Compensation Corporation (ACC), a leader in adopting commercial off-the-shelf and bespoke AI in the public sector, has made its Generative AI Models and Services Policy available proactively. (ACC, 2024)

1. **Transparency** is at the forefront of any Generative AI usage
2. ACC will have **human oversight** included throughout the use of any Generative AI Model
3. Data privacy and security are paramount
4. ACC will actively protect Mātauranga Māori, tikanga, and taonga (**Māori Protected Materials**)
5. We will comply with all applicable **laws and associated policies**
6. All Generative AI Models and Services must have a focus on **ethical use**

7. We will **collaborate with relevant stakeholders** when considering or using Generative AI Models and Services
8. Clarity on usage purposes
9. We will consider and take reasonable steps to protect and respect ACC and third-party **intellectual property** rights

This policy is relatively general compared to Te Whatu Ora's GenAI policy. ACC does not explicitly incorporate domain-specific considerations regarding how GenAI affects ACC's clients or their statutory roles in care, recovery and injury prevention. ACC has also operationalised parts of the NIST AI Risk Framework and has established a Generative AI Advisory Group to provide oversight and expert technical advice to support existing decision-making governance groups (ACC, 2024).

The main novel contribution of ACC's GenAI policy is the active protection of mātauranga Māori. Protection of mātauranga Māori, as required under the Treaty of Waitangi, cannot be guaranteed once processed by GenAI tools: not only because they operate in jurisdictions where Māori control cannot be guaranteed, but also because AI tools themselves pose a risk to the integrity and potential of such material. Like HNZ's framework, ACC's policy reveals further broadly applicable considerations absent in all-of-Government guidance.

4.4. Summary of findings

The analysis in this chapter reveals a guidance ecosystem that has proliferated around a sound legal foundation. System leaders have added layer upon layer as each new technology or political context has demanded a response. The result is an ecosystem that is comprehensive, but practically unwieldy. Practitioners implementing algorithms and AI systems must navigate four standards, five principle-based frameworks, and eight guidance tools, in addition to their fundamental legal obligations.

This proliferation obscures what, in fact, is a coherent underlying legal framework. As Section 4.1 shows, New Zealand's existing legislation already provides meaningful constraints on the design of algorithms and AI systems in the public sector. The OIA's right to reasons under s23 constitutes an effective prohibition on opaque, unexplainable decision-making where individuals are personally affected. The Ombudsman's investigatory powers impose an implicit duty to exclude irrelevant

factors from any automated system. The Public Records Act demands attribution of AI's contribution to any official output. Unlike the aspirational commitments of existing guidance, these are constitutional obligations – enforceable today. The main challenge lies not in what the law actually requires, but how guidance is framed. This implementation gap is the primary challenge I identify for the trustworthy delivery of algorithm and AI systems in New Zealand's public sector.

The most significant issue with the existing guidance is the parallel development of algorithm guidance by StatsNZ and AI guidance by DIA. The Charter and algorithm impact assessment, and the Public Service AI Framework now cover much of the same ground. Table 7 (page 84) demonstrates how nearly every PSAIF principle has a counterpart in the Charter (except for the PSAIF's security principle covering the ground of GCISO guidance). A practitioner within a Charter signatory agency must, in principle, consult both frameworks and reconcile any differences on their own.

Agency-level policies discussed in Section 4.3 illustrate how agencies pick up the responsibility of operationalising guidance and integrating fragmented instruments into a coherent playbook. MSD's Automated Decision-Making Standard and Model Development Lifecycle filled the gap left by an initially vague Algorithm Charter. Health New Zealand published their own AI governance framework centred on trustworthiness and has yet to sign the Algorithm Charter in the four years since its amalgamation. ACC's generative AI policy goes further than either system lead's guidance by meaningfully addressing some of its Treaty obligations to Māori – through the prohibition of surrendering materials to offshore AI systems it cannot protect on behalf of Māori. However, none of these contributions are agency-specific; these are practices applicable to all agencies if system leaders are aware of these gaps.

The path forward I recommend does not require new legislation. The primary challenge is not one of legal coverage but of implementation infrastructure. The guidance that gives operational effect to existing legal obligations is fragmented, unevenly mandated, and variably adopted. The guidance ecosystem merely requires consolidation and elevation: removing duplication, articulating existing legal obligations, and mandating other best practice that is too operational to be legislated. Chapter 5 develops this argument in detail.

5. Re-engineering the government AI guidance ecosystem

This final chapter makes recommendations for solving the challenges with existing guidance instruments. I begin the chapter by arguing that trustworthiness is the correct strategic goal for the development and use of algorithmic and AI systems. Trustworthiness should be at the heart of a new guidance instrument I recommend: the **New Zealand Artificial Intelligence Manual** (NZAIM). Much like the NZISM prescribes both common controls and technology-specific controls, a NZAIM can help to streamline AI practice, much like the NZISM streamlined cybersecurity practice. I propose a taxonomy that practically but adequately specifies the depth of which the NZAIM should prescribe guidance. I also propose consolidating guidance instruments at each level of the hierarchy to make it easier for agencies to implement and for auditors to verify the necessary controls.

These recommendations are informed by four principles derived from software engineering and adapted for application to policymaking. Existing guidance instruments exhibit fragmentation and imprecision; these issues also emerge in software development itself. Both are linguistic artefacts that communicate structured logic and are susceptible to risks of undesirable complexity. Just as principles have emerged to mitigate or prevent the accrual of undesirable technical complexity, so too must guidance instruments be designed to minimise undesirable complexity that gets in the way of development.

A **durable** guidance ecosystem outlasts technological trends and political cycles. Anchoring the ecosystem to enduring commitments reduces the compliance burden that arises from a fragmented guidance ecosystem. System leaders must move on from reactive reinvention of novel principles for novel techniques, tools or policies. A durable guidance ecosystem must also be regularly evaluated to ensure it meets the desired outcomes. Software engineers make a similar call when they incur "technical debt" by opting for quick-and-dirty solutions over maintainable, durable ones (Tom et al., 2013). However, accrued technical debt must be addressed once stable. Chapter 4 shows that the proliferation of different guidance instruments has resulted in similar, undesirable levels of "guidance debt". Durability is also a concern in broader public

policy, particularly, though not exclusively, in relation to technical regulation (Aspray & Doty, 2023).

An **adaptable** guidance ecosystem easily pivots in response to new legislative, political or technological developments. This ecosystem atomises guidance as components maintained by specific expert agencies. New guidance should be designed on top of existing guidance without major modifications. Adaptability is linked to Dijkstra's (1982) principle of separation of concerns, where complexity is decomposed into individual components that deal with only one aspect of the problem. Adaptability is also a key concern within public policy, with strategies like "decompos[ing] complexity into smaller, tangible problems" noted as effective in promoting adaptability (Janssen & Voort, 2016).

A **cohesive** guidance ecosystem identifies universal challenges and provides a unified approach that applies across the broadest range of use cases. Only when unique challenges emerge within a technique or use case is specialisation justified. Cohesion is linked to the concept of abstraction and inheritance in software engineering, where universal concerns are abstracted out at a lower level. Specialised uses do not re-specify; instead, they inherit the established standards of abstracted guidance (GeeksForGeeks, 2025). In public policy, cohesion across applications can be achieved through a principles-based approach (House of Lords, 2019).

An **actionable** guidance ecosystem provides the level of detail required to translate legal obligations and normative commitments into specific, observable technical requirements that practitioners can implement and that independent reviewers can audit. Actionable guidance can follow a design-by-contract ("DbC") approach as in software engineering. DbC centres on a contract outlining how code should behave, embedding correctness into the development process. Actionable AI guidance defines clear requirements for an AI system to serve a legal or normative objective without prescribing how the AI system should work, ensuring the guidance is adaptable for development yet verifiable for audit (Patch, 2015). Actionability has also been identified by AI policy practitioners as a deficit in existing AI ethics principles (Stix, 2021).

My eight recommendations, as summarised in Table 8, thus align with these principles, promoting either durability, adaptability, cohesion or actionability of a future guidance ecosystem:

Table 8: Summary of recommendations of my analysis based on their alignment to engineering principles.

Recommendation	Alignment to principles
5.1. Adopt trustworthiness as the desired strategic outcome across the guidance ecosystem	Improve **durability** by recentring the guidance ecosystem as fostering public trust in government use of algorithms and AI.
5.2. Develop a New Zealand Artificial Intelligence Manual	Improve **cohesion** by developing a New Zealand Artificial Intelligence Manual (NZAIM) integrating all necessary technical guidance in one place.
5.2.1. Provide guidance for different system types	Improve **actionability** by distinguishing between different types of algorithms and AI in the NZAIM while recognising the common challenges between them.
5.2.2. Provide guidance for different system use cases	Improve **durability** by distinguishing between systems used to make OIA-able decisions, inform systemic decisions with evidence, and supporting systems.
5.2.3. Adopt a technical-contextual taxonomy	Improve **actionability** by adopting a technical-contextual taxonomy within the NZAIM, and an eventual government register of algorithm and AI use
5.3. Consolidate data guidance	Improve **adaptability** by unifying GCDS guidance around a strengthened DPUP+ as a lodestar for agencies to foster a trustworthy data use culture
5.4. Consolidate AI use guidance	Improve **adaptability** by streamlining GCDO guidance that better connect its overarching vision to guidance
5.5. Centralise agency algorithm and AI registers	Improve **durability** by using existing data.govt.nz system as a central register for algorithm and AI use

5.1. Adopt trustworthiness as the desired strategic outcome

"Us, trust: a couple things I can't spell without U."

American rapper Big Sean in Bieber et al. (2012)*, presumably on the relationship between an authoritative actor like the government, and a vulnerable party like the citizens it serves, emphasising the latter as the focal point in maintaining an enduring relationship – "us" – built on the evidence of trust by such subjects.*

All the instruments of the guidance ecosystem are centred on different strategic outcomes. Safety and security are promoted by the *New Zealand Information Security Manual* and GCDO AI guidance. Effectiveness was last promoted by the *Principles for safe and effective use of data analytics*. Accountability is promoted by law. Public trust is

promoted by GCDS guidance. Responsibility is promoted by GCDO AI guidance. The **durability** of this ecosystem can be improved by aligning the strategic aims of each of the guidance instruments. While each instrument serves a specific output, they should be framed – at least in high-level aggregate guidance – as promoting the aggregate strategic outcome of promoting public trust through the development of trustworthy systems.

Trust, as summarised by previous government-commissioned research into trustworthy automation (Brainbox Institute, 2021), is "a psychological state where someone is willing to be vulnerable to another's power over them, based on positive expectations of that person's actions". Trustworthiness is measured by public trust, but these are distinct concepts. While trustworthiness promotes enduring public trust, the vulnerable party may mistakenly place trust in the authority due to omission or deception.

Unlike the Algorithm Charter, the Public Service AI guidance did not adopt trust and trustworthiness as its guiding vision, instead calling on agencies to "adopt AI responsibly". However, a government can adopt AI responsibly yet remain untrustworthy, given the ambiguity of the locus of responsibility across different scenarios. This distinction is pertinent in AI system design. For example, an agency may act responsibly by optimising for fiscal efficiency in system design. Such a design is responsible for the state, but it may produce errant, unreliable conclusions of the subject, who thus questions the trustworthiness of the system and of the implementing agency. Centring AI system delivery on trustworthiness has no such ambiguity, as the locus of responsibility is clear. Trustworthiness is the demonstrable state that emerges only from acting responsibly toward the vulnerable subject in this relationship of trust: the citizen. Thus, trustworthiness is a comparatively more **durable** strategic objective than the current Public Service AI Work Programme.

Trustworthiness encompasses all the objectives of each guidance instrument. Trustworthiness can be derived from assurances of safety, effectiveness, and accountability (Brainbox Institute, 2021). Responsibility, by contrast, describes a disposition or process – an input – not a measurable outcome: the willingness to ensure objectives such as safety, effectiveness and accountability. However, these three are merely outputs of a trustworthy design process. For example, Palantir Gotham could be evaluated as the most security-hardened, high-performing, cost-effective, and auditable solution for an integrated data system (such as StatsNZ's IDI

or Te Whatu Ora centralising DHB databases). However, the perception around Palantir, driven by its association with authoritarian practices like predictive policing and aggressive immigration enforcement, creates a significant deficit of perceived appropriateness. Appropriateness is only briefly mentioned in existing all-of-Government guidance, with the introduction of the Algorithm Charter explicitly calling for "the government to use these tools in *appropriate* ways", notably without an explicit commitment in the body stating how. Appropriateness also underpins Health New Zealand's AI governance framework and is always the first question asked of any potential AI use. It is appropriateness, assured through justification of proportionality, that bridges the gap between compliant AI and trustworthy AI.

Public trust and confidence are already measured objectively by most public-facing government agencies through surveys, often serving as a key performance indicator in mandatory annual reporting (e.g. New Zealand Police, Public Service Commission, ACC). This existing quantitative mechanism, which provides a basis for agencies to analyse how their use of AI drives public trust, makes trust and trustworthiness more amenable to consistent evaluation than other strategic objectives, which will differ between agencies based on their operating environments. The **durability** of the ecosystem is improved by anchoring strategic AI guidance to enduring measurable (if not already measured) outcomes.

Trust is also heavily influenced by historical experience. For Māori who historically experienced suppression and oppression by these same government agencies, some argue institutional scepticism is a rational baseline (Gray, 2021). For migrant communities, this institutional scepticism stems from more recent experiences with oppressive governments in their countries of origin (Chen M. , 2019). Demonstrating trustworthiness, not just ambiguous responsibility, is required to overcome such barriers to trust.

Brainbox (2021) also notes that trust is not a universal socio-cultural phenomenon. For example, promoting Māori trust in algorithms and AI will require understanding what is appropriate in the context of Māori values. The GCDS's Ngā Tikanga Paihere has already been developed in consultation with Māori experts to align data practice with tikanga Māori. Proposals for deeper integration of this framework are discussed in Section 5.3.

5.2. Integrate fragmented technical guidance in a New Zealand Artificial Intelligence Manual (NZAIM)

This guidance ecosystem lacks a key instrument that offers an authoritative source for controlling risks and promoting norms identified by other instruments. To enhance cohesion, this unified guidance should organise all objectives across each instrument, mandate or recommend technical controls, and explain their rationale. This approach is already used by the NZISM, giving effect to the Protective Security Requirements by prescribing technical controls. The NZISM offers agencies flexibility in the controls it recommends but does not mandate, only requiring that they document acceptance of the risk of non-use. These controls are supported by a structured certification and accreditation process that provides a standard pathway for system assurance. This approach has standardised cybersecurity practice in government and transparently models best practice for non-government entities. A New Zealand Artificial Intelligence Manual (NZAIM) can not only promote these outcomes for government AI but also for private AI by providing a consumer signal that accredits best practice.

The existing AIA toolkit can serve as the foundation for developing an NZAIM, given that it fulfils a similar purpose under a different framing. The AIA toolkit is merely a methodology for conducting an algorithm-specific risk and impact assessment and subsequent control selection. No mechanism exists to audit (or otherwise peer review) the implementation of AIA-identified controls to the same extent as security and integrity controls identified in NZISM processes. Thus, an NZAIM would augment the existing security-focused C&A process by also validating a system's correctness, lawfulness and fairness as identified by an AIA.

Nevertheless, this redevelopment presents an opportunity to review the AIA toolkit, as no such review has been signalled despite it being two years old. This oversight is surprising given the advances in technology, AI policy within the public service, and the public service's uptake of AI. However, a potential explanation is the inopportune timing of the AIA toolkit's release: at the inflection point of a change of government, which may have disrupted the incoming administration's engagement with detailed policy continuity work.

A review of the AIA toolkit should assess its effectiveness among end users. There are currently no published AIAs by any agency, despite many agencies

deploying AI tools with only PIAs, not even the basic algorithm threshold assessment. Such a review may assess the effectiveness of communication and engagement around this tool, as well as the tool's usability. Usability may be improved by streamlining the assessment. 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. This alignment should improve the AIA's **cohesion**.

A review may also consider how to incorporate the Māori AI Governance model, not just through the Partnership with Māori commitment, but throughout the entire toolkit. For example, Question 6.3 (Storage) queries the location of data storage in relation to Privacy Act obligations, but should also query whether there are significant interests by iwi Māori to retain data onshore to enhance their sovereignty. Notably, no specific Māori data experts were acknowledged as having provided feedback to the toolkit.

Delegating technical guidance to the NZAIM will enhance its **durability**, allowing the AIA to focus on impact identification. For example, MSD's Model Development Lifecycle is mentioned throughout the user guide, which it relies on to supplement 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.

5.2.1. The NZAIM should provide guidance for different system types

Chapter 2 reveals both the unique risks that each type of algorithm and AI system faces, as well as common risks. Common risks can be managed by common controls, such as documentation to promote awareness and monitoring. Guidance can be more **cohesive** by identifying these technical commonalities.

Another set of common risks is managed by different controls. For example, the risk of concept drift and data drift across all techniques is stemmed by regular monitoring and evaluation. However, evaluation methods can vary significantly across different categories of algorithms and AI. For supervised learning, quantitative evaluation is built into the model development process. For predefined algorithms, quantitative testing is achieved through traditional software testing frameworks.

However, unsupervised learning and generative AI require manual post-hoc evaluation. Generative AI evaluation is much more intractable given its freeform, probabilistic nature. Given that an NZAIM audit will centre on such evaluation, the NZAIM will be more **actionable** if it specifies standardised evaluation methodologies.

Another set of risks is unique to each evaluation paradigm. For example, the deep and freeform nature of GenAI leaves it vulnerable to a much wider range of risks, as explained in Section 2.5.4. The mathematical optimisation required for GDO makes it uniquely vulnerable to poor outcomes due to a weak problem definition within the constraints of quantitative measurement.

The NZAIM should retain scope over algorithms, in addition to AI, given the common controls across both. This distinction is not evident in existing AI-specific guidance, and 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, Dutch childcare benefits scandal) 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, whether it is a traditional algorithm or an AI system.

A similar risk arises when conflating AI with GenAI. AI systems have long been deployed across government using traditional predictive techniques. 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.

Therefore, a technical categorisation (as visualised in Figure 14) can improve cohesion by eliminating terminology that is not useful for monitoring and governance. This categorisation eschews the broad, polysemous term "AI" because the other categories collectively encompass these AI techniques. This categorisation also avoids using architectural descriptors such as "deep learning" (which describes a variety of ML applications), "agentic" (which can be evaluated as a handcrafted algorithm, a standard LLM, or a goal-driven optimiser), and "reinforcement learning" and "evolutionary computation" (both of which are goal-driven optimisers).

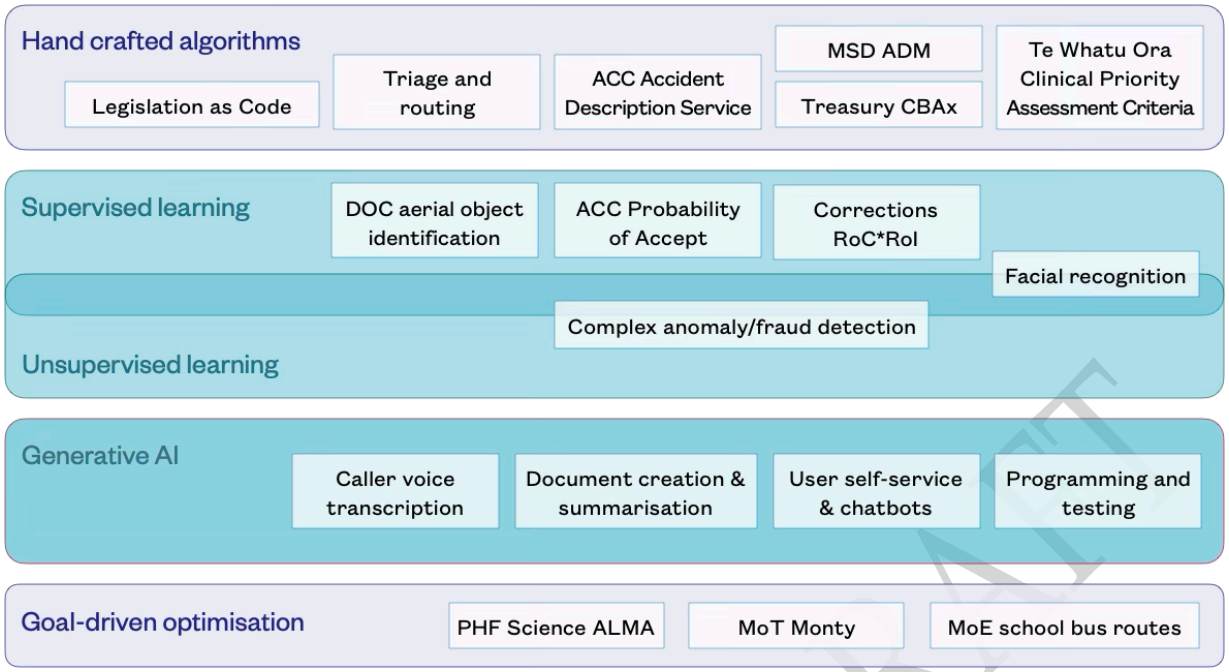

Figure 14: Taxonomy of system types based on major evaluation paradigms, with mentioned examples grouped.

5.2.2. The NZAIM should provide guidance for different system use cases

Given that relevant legislation is technology-agnostic yet context-specific, the guidance ecosystem must also consider the unique legal obligations for certain algorithmic and AI use cases. Context-aware guidance will increase its **durability,** providing baseline obligations for a particular use case regardless of the technology used, enabling agencies to safely innovate with new technologies by anticipating requirements without waiting for explicit guidance.

All-of-Government stocktakes (StatsNZ, 2018) and (DIA, 2024) used different ways of categorising algorithmic and AI use cases. The StatsNZ categorisation was at a higher level, aligned with the separation of service delivery and policy functions, with a category for internal rules. This higher-level categorisation is more durable than DIA's, which grouped use cases arbitrarily based on common tactical benefits, such as "boosting productivity and efficiency" and "enhancing customer experience". This categorisation was fit for purpose in the 2024 stocktake, given that it mainly served to stimulate discussion within agencies in the early stages of implementing AI.

To improve the durability of StatsNZ's (2018) categorisation, my modifications ground it in the unique legal obligations and risk profile, while accommodating new

technologies that have emerged since then. Such a categorisation must be broad enough to cover as many possible applications of techniques as possible, even as new innovations emerge. I reorganise the existing categorisation below, and visualise it in Figure 15:

- Non-conclusive business rules and AI systems that automate or assist core, impactful job functions now belong to the **administrative use case** category.
- Operational algorithms and conclusive business rules now belong to the **front-line use case** category, defined as systems subject to the OIA rights to reasons and rules. Additionally, generative AI uses that reach their own independent conclusions are included under front-line.
- Algorithms for policy development and research now include operational research techniques under the **research use case** category.
- Algorithms that otherwise do not automate core, impactful job functions, make disclosable decisions, or produce primary evidence are not subject to additional scrutiny beyond that already required by existing technology, architecture, and security governance.

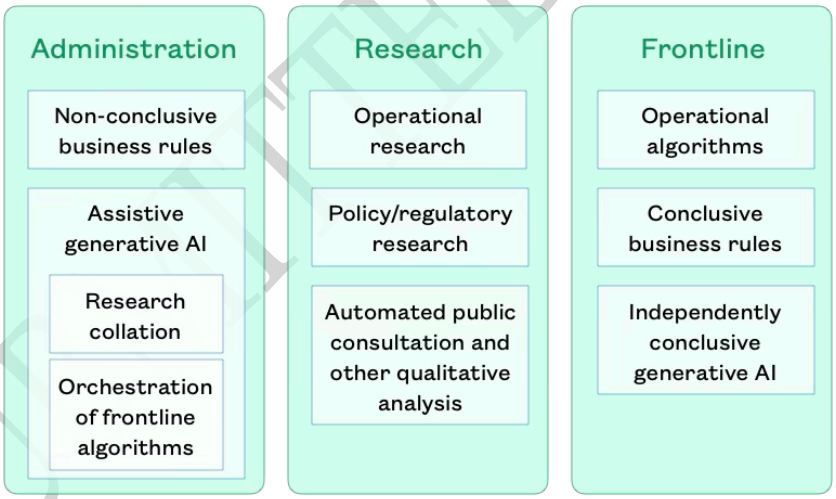

Figure 15: Taxonomy of use cases based on the relevant legal obligations within each category.

Most business rules and generative AI tools are **administrative systems**. These systems do not directly lead to a conclusive decision, but they still have a *material impact* on government operations by automating *core* activities. The distinction between an administrative system and a non-core, non-impactful system is ultimately the judgment call of an agency with its unique risk appetite. These systems may use personal data, subjecting it to the provisions of the Privacy Act. However, these systems are not subject to the same level of transparency and accountability as other

use cases. Nevertheless, these systems must still be assured and continuously maintained as safe, effective and appropriate. These systems are still subject to the Public Records Act, which requires recordkeeping for both the system itself and the attribution of its outputs (and their downstream use).

Front-line systems include algorithms, AI, and GenAI systems that directly and independently make decisions or recommendations regarding an individual that impact their rights or interests. This category is subject to the highest level of transparency and accountability as, by definition, all systems in this category are subject to the right to rules and right to reasons under the OIA. Only GenAI systems that make independent judgement calls (for example, risk profiling based on unstructured client documents and notes) are included under frontline. GenAI systems that follow a prescribed structure for making decisions and recommendations can be considered administrative – the prescribed algorithm will be the primary subject of an OIA request.

Guidance for front-line systems should still be followed even if the decisions made by these systems are within the scope of a rules or reasons request, but may be withheld under the refusal grounds available in such a request. For instance, while a request for a border agency's facial recognition rules may be withheld on the grounds that it risks prejudicing New Zealand's security, it would nevertheless benefit from the considerations (such as ensuring fairness with respect to protected attributes) applied to other front-line systems, as the high-risk nature of its decisions remains identical.

Algorithms for policy development and research should also include operational research employing algorithms and AI, under the **"research"** category. This group includes both operational and policy research techniques that generate data to form the evidence base for system-level decision-making. This category remediates an oversight StatsNZ made in its algorithm stocktake. They stated that this category has "no direct or significant impact on individuals or groups." While this focus was understandable at the formative stage of all-of-government algorithm governance, this assertion is inaccurate as the commission or omission (i.e. opportunity cost) of policy substantially impacts vast swathes of individuals and groups. Those affected include those who do not rely on specific government interactions, unlike operational algorithms, thus arguably making their impact relatively greater. This impact, however, is not directly perceived, as they never

interact directly with the government when such decisions are made and thus are not subject to OIA rights.

While it can be argued that policy accountability rests at the ballot box, the kinds of policy decisions supported by analytical research by government officials are usually not the ones on which the government is popularly elected. Policies that gain visibility during an election are often implemented by virtue of the electoral mandate obtained from campaigning on those policies, regardless of whether they are based on empirical evidence[2]. Policy initiatives that originate with officials seldom attract comparable public visibility, often appearing procedural and esoteric, making them harder to hold democratically accountable. Ministers implicitly trust policymakers (who should be experts in their domains) that the evidence base they act on is robust and accurate, to ensure that policy changes are beneficial. Therefore, it is equally vital to ensure – as in the operational space where algorithms may enhance decision-making accuracy but could introduce new or existing biases and harm – that algorithms for policy development are promoted to improve predictive accuracy while mitigating their potential harms.

Suboptimal policy modelling may result in the adoption of suboptimal or even harmful rules and interventions. Policy modelling and data that presage the withdrawal of interventions may deprive individuals and communities of services that truly work and deliver the right outcomes for them. Additionally, government ministers may lose confidence in their agencies when poor or untimely modelling and data lead to suboptimal ministerial decisions that attract public scrutiny. Past examples of arguably policy modelling that led to suboptimal policy decisions include the Ministry of Education's 2023 Teacher Demand and Supply Planning Projection (Walters, 2025) and the Reserve Bank's under-forecasting inflation since the June 2021 quarter, which all forecasts of the major banks also missed (Bohm & Sing, 2022).

[2] cf. the Sixth National Government who campaigned on the three strikes policy, which officials evaluated as having "limited evidence that it reduced serious crime" when it was last implemented (Ministry of Justice, 2024); or the Sixth Labour Government who campaigned on fair pay agreements, which officials assessed as not being founded on empirical evidence (Treasury, 2018).

At present, policymakers employ simple analytical methods that (as discussed in Section 2.2.2) are often limited in accuracy because they do not adequately capture the complexity of patterns and relationships within datasets, particularly those measuring intricate systems with a vast array of influencing factors, such as climate or the economy. Methods like ML (as outlined in Section 2.4) can yield more accurate, timely models that help policymakers make decisions more confidently and rapidly.

5.2.3. Adopt an intersectional taxonomy based on the system's technique and use case

Overlaying the technical and use considerations for trustworthy AI system delivery is critical to develop **actionable** guidance. A two-way technical-contextual taxonomy, as illustrated in Figure 16, should provide sufficient practical specificity for monitoring and governance in the New Zealand context.

This intersectional approach provides a more structured method of impact assessment and management, demonstrating how effects can manifest differently based on where and how an AI system is implemented. For example, the risk of automation bias can manifest differently in different intersections of frontline, research, supervised and generative systems:

- For frontline supervised AI systems, the risk manifests as confirmation bias. The systems' discrete, unnuanced predictions can presuppose the evidence sought in further human-led investigation or validation designed to mitigate automation bias.
- For frontline generative AI systems, the risk manifests as fluency bias (Ghafour, 2025). As generative AI systems are trained to produce statistically likely content, they not only risk making erroneous decisions but are also trained to frame these decisions in a fluent, authoritative manner, increasing the likelihood of succumbing to automation bias. Reasoning models exacerbate this risk by producing fluent, authoritative reasoning that misrepresents the true reasons for a prediction (Turpin et al., 2023).
- For research supervised systems, the risk manifests as quantification bias (Chang et al., 2024). Here, the output forecasts or analyses are given increased credence over the qualitative intuition simply by virtue of being measurable.

- For research generative AI systems, where generative AI devises a plan or develops the code that synthesises evidence, the risk manifests not just in the output as fluency bias as previously mentioned, but also a bias towards assessing the outputs without assessing the methodology that led to the output. Faulty assumptions or logic, or faulty implementation of the methodology, can lead to faulty outputs.

Figure 16 illustrates how existing and potential use cases can be more intuitively organised under this taxonomy. This taxonomy can also serve as the basis for a visual all-of-Government register, as described in Section 5.5. Figure 17 also shows how the taxonomy highlights the gaps in existing guidance by mapping instruments to the areas they were designed to influence.

5.3. Consolidate data use guidance around a mandatory DPUP standard

In software engineering, when several pieces of code perform the same job in slightly different ways, they can be cleaned up and consolidated into a single version. This process, known as refactoring, makes the code easier to read and understand by removing unnecessary clutter. Refactoring also enhances maintainability, as improvements only need to be made once in the consolidated version to apply everywhere.

Both the Charter and the DPUP do the same job for slightly different purposes. Each articulates expectations around the use of data or algorithms that extend beyond existing obligations, with the aim of promoting public trust through the respectful and safe use of data and algorithms. However, each falls 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.

Figure 16: Map of existing algorithms and AI uses in the New Zealand Government under the technical-contextual taxonomy.

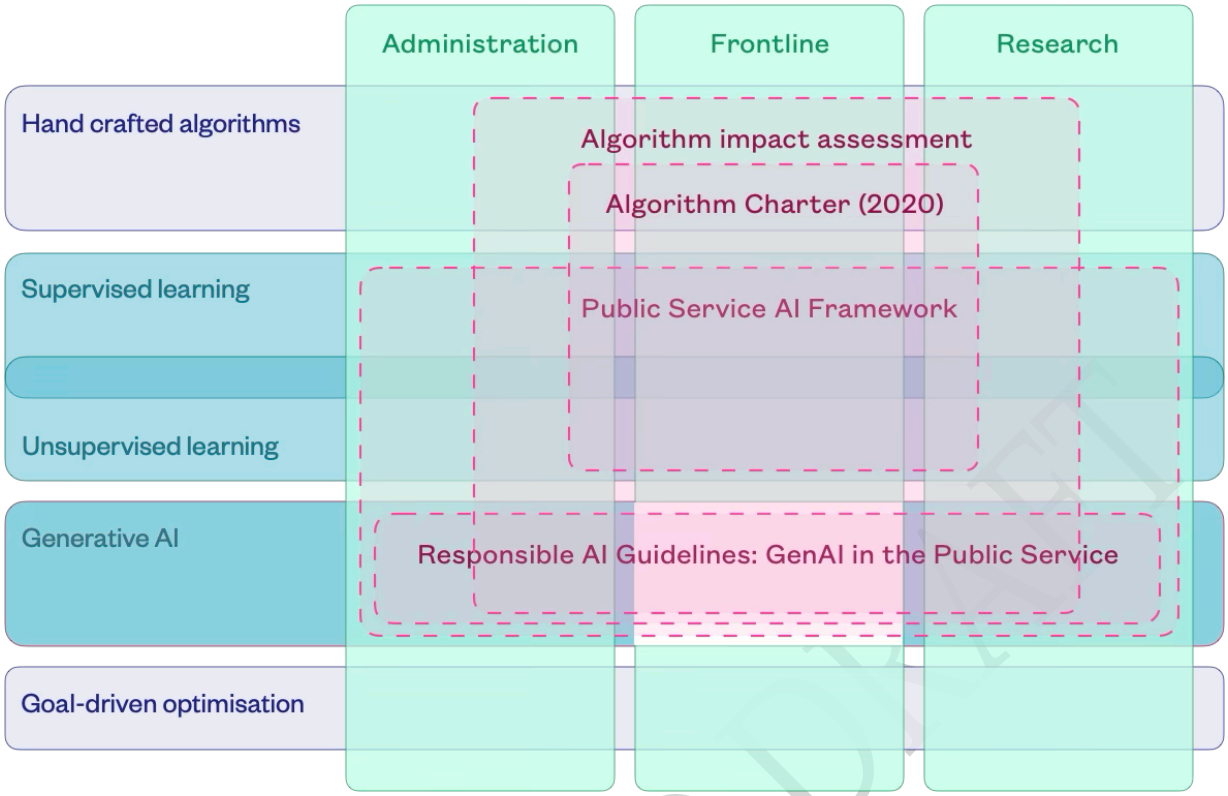

Figure 17: Map of existing guidance instruments mapped to the techniques and use cases it was designed to influence.

By recognising their similarity, the data guidance ecosystem can become more **adaptable** by reframing 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 much 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 emphasises the fundamental, foundational nature of GCDS guidance. Trustworthy use of any information technology does not happen without the trustworthy use of the information that communities entrust to agencies. Conversely, the comprehensive trustworthy use of this information begets the trustworthy use of any novel information technology, provided the right engagement mechanisms are already in place. While technical practitioners play a critical role in implementing controls, they do not typically possess the institutional

authority or capability to set cultural expectations for respectful data practices and to sustain active, meaningful community engagement. These are inherently organisational questions that require explicit whole-of-agency commitment and support. To be effective, this commitment should stem from unified departmental efforts and policies, 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.

Consolidating GCDS policy also creates an opportunity to address other critical actions the GCDS has yet to take:

- Promoting Ngā Tikanga Paihere as general advice for culturally appropriate use of Māori data across 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, and previous technology-specific research by Toi Āria and Te Kāhui Raraunga.
- Assessing compliance with GCDS guidance 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 whether, the GCDS at present benchmarks organisational data maturity. Their Data Maturity Assessment was last mentioned as being in the pilot stage in early 2023, prior to StatsNZ disestablishing its Data System Leadership branch.

After streamlining normative commitments, what remains of the Algorithm Charter consists of operational obligations already mandated by legislation and standards. These obligations are best clarified in the NZAIM.

Streamlining normative commitments would also present an opportunity to address a potential deficit that the legal analysis alone does not capture: further simplifying and optimising public communication around what safeguards the Government already follows. The instruments reviewed in Chapter 4 are, without exception, addressed to agencies and practitioners – to the people responsible for building and deploying algorithmic systems, not to the people those systems act

upon. The Algorithm Charter is published deep within a satellite website outside an agency's core information architecture; the AIA toolkit is a practitioner methodology. The OIA's right to reasons is only exercisable by someone who knows it exists; Ombudsman complaints require an awareness that a complaint is possible. These safeguards are not deliberately communicated to the individuals whose livelihoods are affected by these systems. The benefit of GCDS guidance is that it is already articulated in plain language at the highest level, unlike the practitioner-focused GCDO guidance. The issue with GCDS guidance is that multiple different guidance instruments provide slightly different commitments. Plain, normative commitments to trustworthy data use can improve the Government's messaging. But for such commitments to carry weight, they must also manifest in actionable guidance developed by other actors in the guidance ecosystem.

5.4. Consolidate AI end-use guidance around the PSAIF principles

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. If viewed as a strategy, the PSAIF is fit for purpose as it provides a strategic direction and defines what good looks like. While it helps agencies devise their own AI strategies, it does not define how they execute those strategies at the operational level. 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 imposes significant constraints on the use of AI.

Instead, the Responsible AI Guidance (RAIG) – a suite of guidance with the first of the series focused on Public Service GenAI (RAIG-PSG) – defines the 'how'. It shows promise as a comprehensive, unified AI guidance framework. Its GenAI guidance offers useful, novel advice on that technology and references existing guidance, such as the NZISM and GWS. However, RAIG-PSG does not sufficiently integrate best practice from the prior Algorithm Charter or the subsequent RAIG for businesses. Furthermore, the RAIG-PSG refers to, but does not structure its guidance around, the PSAIF principles.

Section 5.3 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.

MBIE, despite holding no system leadership over public sector AI, has produced more detailed technical guidance than the GCDO — a gap that reflects the siloed nature of the current ecosystem. As the leader of microeconomic policy, MBIE has been tasked by Cabinet to help businesses use AI responsibly. This work has culminated in the National AI Strategy and the Responsible AI Guidance for Businesses (RAIG-B), released in July 2025. This specialisation of MBIE and GCDO guidance has not produced markedly different guidance between them, despite the difference in risk profiles and incentives between the public and private sectors. On the contrary, RAIG-B offers useful technical guidance applicable to the public sector and could be adopted in a more 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 on recognising that bias and unfairness stem from poor-quality data, resulting in poorly performing AI. RAIG-PSG treats this as an independent concept, mitigated only by process controls. RAIG-B correctly points out that technical controls are just as effective, offering useful examples like "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".
- Prompting the consideration of the legality and ethics of certain collection and use of data. While RAIG-PSG acknowledges agencies' Privacy Act obligations, it does not address other types of legally and ethically contentious data collection identified in RAIG-B, such as using copyrighted work without permission and web scraping. RAIG-B emphasises that there are options available for procuring ethically trained AI systems. RAIG-B advice around Māori data is more comprehensive and affirmative of Māori sovereignty than Crown guidance, despite only the Crown having formal fiduciary duties to iwi Māori.

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 is not organised around the points identified in the PSAIF. Furthermore,

the RAIG-PSG has opted to use the OECD AI Principles, rather than the principles of the PSAIF, which have been "inform[ed]" by, but are simpler than, the OECD principles. The PSAIF prompts 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 and the anti-discrimination obligations under the Human Rights Act, and it references the Privacy Act only 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.

Two further structural improvements to the GCDO guidance ecosystem would improve its cohesion:

- 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 by establishing precedents through which the PSAIF can be given practical effect, rather than agencies relying on prescriptive guidance that allows each to use its own structure.

5.5. Centralise agency algorithm and AI registers

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 (n.d.). 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 by system lead agencies for

StatsNZ (2018) and DIA (DIA, 2024). 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 functions as a low-effort mechanism of implicit accountability in the absence of legal enforcement. An all-of-government register has been advocated for since Gavaghan et al. (2019) and the Taylor Fry (2021) evaluation. This register can promote the dissemination of best practice in addressing common challenges by disclosing impact controls applied in such cases. Such a register also creates a soft compliance dynamic, with agencies that do not report implicitly signalling lower governance maturity, thereby incentivising remediation as a form of reputational risk management. Comprehensive reporting on algorithm and AI use can also foster public trust in an agency's use by clearly delineating where AI is and is not used. Importantly, such a register can be created without exposing sensitive details. At a minimum, a register should disclose:

- Basic context around when, where, and for what purpose the algorithm is used
- Answers to the algorithm threshold assessment's questions
- If ATA's threshold is met, when and who (job title or qualification sufficient) conducted the last regular peer review to assess for unintended consequences, and how they acted on this information – or justify why a peer review has not been recently conducted (e.g. static legislation and environment minimises risk of concept drift)

In this regime, sensitive information that may have been justifiably withheld under the OIA still need not be proactively disclosed. Nevertheless, 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 provide assurance regarding the lawfulness of decision-making subject to judicial review, in accordance with its establishing legislation and the previously mentioned wider legislative framework. A peer review may also scrutinise alignment with the Charter's non-mandatory commitments. The AIA toolkit already provides a suitably comprehensive framework to structure such a peer review.

In Canada, the DADM mandates transparency by requiring the publication of algorithmic impact assessments on the federal Open Government Portal prior to the deployment of such systems. This website uses CKAN, which is also the backend for

New Zealand's data.govt.nz dataset catalogue, maintained by the GCDO. It is trivial to add an "Algorithm" or "AIA" tag within data.govt.nz to replicate Canada's ADM register. Given that agencies in New Zealand already have established procedures and APIs for uploading content to this portal, agencies will not experience a steep learning and implementation curve.

5.6. Target model for a future guidance ecosystem

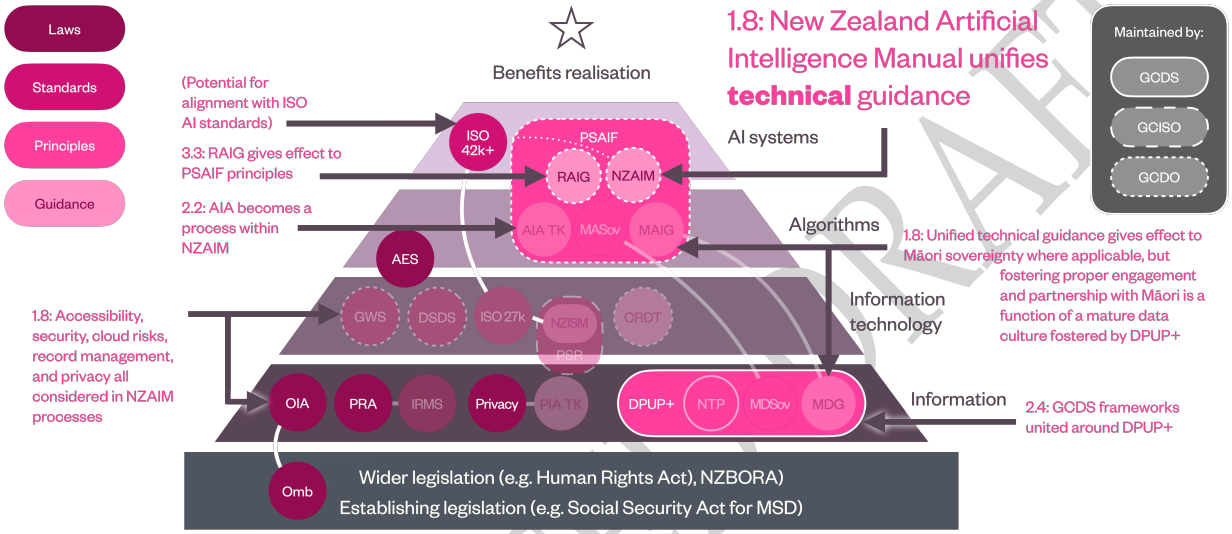

Figure 18: Re-engineered guidance ecosystem with consolidated data use policies and algorithm/AI policies, with the NZAIM highlighted as a new, unified technical guidance instrument.

As this thesis argues that the existing legislative framework already provides the basis for compelling best practice where needed, this chapter's recommendations focus on re-engineering supporting frameworks and guidance. Figure 18 summarises the proposed state of a more streamlined guidance ecosystem.

Normative commitments around the trustworthy, considerate use of any information, particularly system-specific normative commitments found in the Algorithm Charter and Ngā Tikanga Paihere, and partnership-specific commitments promoted by Te Kāhui Raraunga, can be refactored around the existing Data Protection and Use Policy (DPUP). A single, strengthened, mandated DPUP+ will act as both the blueprint for agencies to build a high-performing, trustworthy, and considerate data culture and a benchmark to independently assess an agency's performance against best practice.

Operational considerations around the development and end use of AI systems can be unified around a New Zealand Artificial Intelligence Manual and streamlined

Responsible AI Guidance instalments, respectively. Both should be organised under the principles of the Public Service AI Framework to clearly illustrate how this strategy is operationalised. An NZAIM will integrate (instead of simply referencing) existing obligations and guidance around accessibility, security, cloud risks, record management and privacy.

This streamlined ecosystem untangles the overlapping roles of the GCDO and the GCDS by delineating – but functionally linking – the ethical "why" and the operational "how". Abstracting the "why" at the lowest fundamental level provides an enduring foundation for any "how", both now and in the future.

6. Conclusion

The New Zealand Government has a long history of exercising power through algorithms and AI. From the RoC*RoI model to ChatGPT-led statutory consultation, the Government has recognised for decades that the delegation of impactful judgement to automated systems carries both significant promise and significant risk. What is new, over recent years, is the scale and scope of this delegation. Recent technological innovations continue to unlock new possibilities, and new harms, in the administration of the state. This expansion is outpacing the development of the conceptual vocabulary of governance frameworks designed to constrain it. This thesis has examined this gap: between what systems can do, how these systems are currently used, what the law requires, what the guidance says, and what agencies do. This thesis proposes a re-engineered guidance ecosystem to adequately close this gap.

The primary argument of this thesis is that this implementation gap is not primarily due to a legislative deficit. The analysis in Chapter 4 demonstrates that New Zealand's existing statutory framework already provides meaningful, enforceable constraints on the design of algorithms and AI in government. Transparency rights under the *Official Information Act 1982*, privacy rights under the *Privacy Act 2020*, recordkeeping obligations under the *Public Records Act 2005*, and accountability mechanisms available through enabling legislation or the Ombudsman collectively constitute a coherent theoretical legal foundation for the delivery of trustworthy AI systems.

The issue is that the guidance ecosystem built on top of this legal framework has grown reactively and without coordination. The result is a proliferation of instruments that overlap in scope, diverge in terminology, and collectively present legally mandated obligations as optional best practice. Agencies navigating this ecosystem must integrate guidance that was never deliberately designed to be cohesive. Agencies may lack the capacity to do so rigorously, which can have a chilling effect on practitioners seeking to unlock the potential benefits these systems may bring or may result in the proliferation of shadow AI systems outside the purview of such governance mechanisms.

The technical and use case reviews in Chapters 2 and 3 reveal two further governance gaps that existing instruments have not adequately addressed. First, goal-

driven optimisation techniques, such as agent-based modelling or evolutionary computation, are already in use across the public sector. However, guidance frameworks do not consider these techniques despite their conceptual similarity to other techniques. Second, the distinction between frontline systems and research systems has been historically underweighted. Evidence-generating systems carry risks comparable in magnitude to front-line systems, yet attract significantly less governance scrutiny. The taxonomy proposed in Chapter 5 addresses both omissions by organising guidance obligations around the actual legal exposure of each system type and use case, rather than around the technical architecture of the system itself.

My recommendations in Chapter 5 do not require new legislation. Instead, they require the deliberate rationalisation of what already exists. The guidance ecosystem must be deliberately engineered based on the *Public Service AI Framework* and must better cohere with the *OECD AI Principles*. Normative commitments can be consolidated around a strengthened, mandatory *Data Protection and Use Policy* that endures agnostically of technology and gives effect to the Crown's obligations under Te Tiriti o Waitangi. Technical guidance can be integrated into a *New Zealand Artificial Intelligence Manual*, modelled and built on the success of the *New Zealand Information Security Manual*. This NZAIM does not introduce new obligations but assembles existing ones in a form that practitioners can easily implement, and auditors can easily verify. The strategic refocusing of all these instruments around trustworthiness provides a durable anchor and a measurable barometer that the ecosystem currently lacks.

These recommendations are shaped by the unique characteristics of New Zealand's public service: a small but decentralised bureaucracy with increasing functional system leadership, a unique constitutional relationship with iwi Māori, and a regulatory tradition that favours principles over prescription. New Zealand is a relative outlier in the extent to which it fragmented its public service as part of the global wave of neoliberal reforms in the late 20th century. Consequently, other countries face a less acute version of the coordination issues this thesis addresses. While system leadership over technology and data in other governments may not be as siloed as New Zealand, the need for trustworthy data governance to precede and underpin any technology governance remains relevant.

These recommendations are also fundamentally theoretical, derived from the literature review and critical analysis. The gaps my thesis identifies were not derived

from qualitative measurement of how guidance instruments have shaped the behaviour of agencies of different sizes and maturities. Empirical research and evaluation into how agencies navigate the existing ecosystem, where further guidance may be needed, and what interventions have demonstrably shifted practice, must be undertaken by system leaders prior to the proposed consolidation to determine what will truly work in practice.

New Zealand has the statutory architecture, the institutional experience, and the right guidance infrastructure. What has been missing is the coordination across siloed system leaders to sufficiently demonstrate to the public that the Government is doing right by them. The Wanganui Computer generated public resistance not because it lacked governance processes, but because it lacked meaningful engagement with the communities it integrated data about, and the trust that such engagement would have established. Fifty years later, we risk the same issue: a governance apparatus that is legible to experts but invisible to the people most affected by its operation. Closing this gap, by enforcing guidance centred on trustworthiness, is not just a matter of parsimonious engineering. In a context of declining public trust in government and rising concern about AI, it is necessary to better sustain the legitimacy of a technologically assisted state and to prevent another eruption of distrust like the Wanganui Computer bombing.

7. Bibliography

ACC. (2018, August 21). *Statistical models to improve ACC claims approval and registration process.* Retrieved from ACC: https://www.acc.co.nz/assets/im-injured/ef79338f63/claims-approval-technical-summary.pdf

ACC. (2024, February). *ACC Privacy Impact Assessment (PIA) - Microsoft 365 Copilot.* Retrieved from ACC: https://www.acc.co.nz/assets/business/Privacy-Impact-Assessment-M365-Copilot.pdf

ACC. (2024, March 15). *Use of Generative Artificial Intelligence (Gen AI) - Our research.* Retrieved from ACC: https://www.acc.co.nz/about-us/research#use-of-generative-artificial-intelligence-gen-ai

Acumen. (2025, March). *Acumen Edelman Trust Barometer 2025.* Retrieved from Acumen: https://acumennz.com/acumen-edelman-trust-barometer/acumen-edelman-trust-barometer-2025/

Albada, M. (2025). *Building Applications with AI Agents.* O'Reilly Media, Inc.

Anthropic. (2025, April 3). *Reasoning Models Don't Always Say What They Think.* Retrieved from Anthropic: https://www.anthropic.com/research/reasoning-models-dont-say-think

Archives New Zealand. (2023, December). *Artificial intelligence and public and local authority records* . Retrieved from Archives New Zealand: https://www.archives.govt.nz/manage-information/how-to-manage-your-information/implementation/artificial-intelligence-and-public-records

Arup. (2020, December). *Building an Agent Based Model for New Zealand.* Retrieved from Ministry of Transport: https://www.transport.govt.nz/assets/Uploads/EvaluationWorkProgramme_Part1.pdf

Aspray, W., & Doty, P. (2023, May 8). Does technology really outpace policy, and does it matter? A primer for technical experts and others. *Journal of the Association for Information Science and Technology, 74*(8), 885-904.

Aurecon. (2024, March). *Enabling digital technologies for New Zealand's circular and bioeconomy, including the role of digital twins.* Retrieved from MBIE:

https://www.mbie.govt.nz/dmsdocument/28289-digital-technologies-digital-twins-and-the-circular-and-bioeconomy

Baker Wilson, K. (2026, January 15). *Cyber-security expert launches petition to Parliament calling for harsher penalties for privacy breaches* . Retrieved from RNZ: https://www.rnz.co.nz/news/political/584086/cyber-security-expert-launches-petition-to-parliament-calling-for-harsher-penalties-for-privacy-breaches

Bengesi, S., El-Sayed, H., Sarker, M., Houkpati, Y., Irungu, J., & Oladunni, T. (2023). Advancements in Generative AI: A Comprehensive Review of GANs, GPT, Autoencoders, Diffusion Model, and Transformers. *IEEE Access*, 69812-69837.

Berryman, J., & Ziegler, A. (2024). *Prompt Engineering for LLMs.* O'Reilly Media.

Bieber, J., Lindal, A., Atweh, N., Jerkins, R., & Anderson, S. (2012). As Long As You Love Me [Recorded by J. Bieber, & S. M. Anderson]. On *Believe*. R. Jerkins, & A. Lindal.

Bloomfield, A. (2023). Developing Future Public Service Leaders for Aotearoa New Zealand. *Policy Quarterly, 19*(1), 3-9.

Bohm, T., & Sing, M. (2022, November). *Evaluating the Reserve Bank's Forecasting Performance.* Retrieved from Reserve Bank of New Zealand: https://www.rbnz.govt.nz/-/media/project/sites/rbnz/files/publications/bulletins/2022/bulletin---evaluating-the-reserve-banks-forecasting-performance.pdf

Boshier, P. (2025). *The Chief Ombudsman's Reflections on the Official Information Act.* Wellington: Office of the Ombudsman.

Brainbox Institute. (2021, April 16). *Research insights - Digital Council of Aotearoa New Zealand: trust and automated decision-making.* Retrieved from Digital Council for Aotearoa: https://ndhadeliver.natlib.govt.nz/webarchive/20230516040816/https://digitalcouncil.govt.nz/advice/reports/towards-trustworthy-and-trusted-automated-decision-making-in-aotearoa/

Brown, P., Wilson, D., West, K., Escott, K.-R., Basabas, K., Ritchie, B., . . . Keegan, T. (2024). Māori Algorithmic Sovereignty: Idea, Principles, and Use. *Data Science Journal*, 15.

Chen, A. (2022). The Algorithm Charter / He Tūtohi Hātepe mō Aotearoa. In A. Pendergast, & K. Pendergast, *More Zeros and Ones: Digital Technology, Maintenance and Equity in Aotearoa New Zealand* (pp. 135-150). Wellington: Bridget Williams Books .

Chen, M. (2019, October). *National Culture and its Impact on Workplace Health and Safety and Injury Prevention for Employers and Workers.* Retrieved from Superdiversity Institute for Law, Policy and Business: https://www.maichen.nz/wp-content/uploads/superdiversity-acc-report-2019.pdf

Chopra, A. (2024). *Large Population Models: Technical Contributions and Open Problems.* Retrieved from MIT Media Lab: http://lpm.media.mit.edu/technical.pdf

Cormen, T. H., Leiserson, C. E., Rivest, R. L., & Stein, C. (2001). *Introduction to Algorithms.* Cambridge, MA: MIT Press.

Courtney, B. (2021). *Unchartered Waters: The insufficiencies of New Zealand's 'world-first' Algorithm Charter in governance of automated decision-making.* Wellington: Te Herenga Waka—Victoria University of Wellington.

Criado-Perez, C. (2019). Chapter 8: One-size-fits-men. In C. Criado-Perez, *Invisible women: exposing data bias in a world designed for men* (pp. 157-168). London: Chatto & Windus.

Croft, J. (2024, January 26). *Identifying drift in ML models: Best practices for generating consistent, reliable responses* . Retrieved from FastTrack for Azure: https://techcommunity.microsoft.com/blog/fasttrackforazureblog/identifying-drift-in-ml-models-best-practices-for-generating-consistent-reliable/4040531

Crown Law Office. (2019). *Te Pouārahi | The Judge over your Shoulder.* Retrieved from Crown Law: https://www.crownlaw.govt.nz/assets/Uploads/JOYS-for-web.pdf

Daalder, M. (2025, September 23). *Health NZ scraps in-house AI tool in favour of private sector.* Retrieved from Newsroom:

https://newsroom.co.nz/2025/09/23/health-nz-scraps-in-house-ai-tool-in-favour-of-private-sector/

Daalder, M. (2026, March 6). *'I want to sounds smart and inspirational': How ministers use AI.* Retrieved from Newsroom: https://newsroom.co.nz/2026/03/06/how-ministers-use-ai/

Department of Corrections. (2007, September). *Over-representation of Māori in the criminal justice system - An exploratory report.* Retrieved from Department of Corrections: https://www.corrections.govt.nz/resources/research/over-representation-of-maori-in-the-criminal-justice-system

DIA. (2024, July 29). *Public Service Al use and options for accelerating Al innovation.* Retrieved from Department of Internal Affairs: https://www.dia.govt.nz/diawebsite.nsf/Files/Proactive-Releases-2024-25/%24file/Proactive-release-of-material-relating-to-Artificial-Intelligence-in-the-month-of-July.pdf

DIA. (2025, January 29). *Public Service AI Framework.* Retrieved from digital.govt.nz: https://www.digital.govt.nz/standards-and-guidance/technology-and-architecture/artificial-intelligence/public-service-artificial-intelligence-framework

DIA. (2026, January 16). *Public Service AI Work Programme*. Retrieved from Digital.govt.nz: https://www.digital.govt.nz/standards-and-guidance/technology-and-architecture/artificial-intelligence/public-service-ai-work-programme

DIA. (n.d.). *How to carry out a privacy impact assessment on your dataset*. Retrieved from data.govt.nz: https://data.govt.nz/toolkit/privacy-and-security/how-to-carry-out-a-privacy-impact-assessment-on-your-dataset

Digital Council for Aotearoa. (2021, April 16). *Towards trustworthy and trusted automated decision-making in Aotearoa* . Retrieved from National Library: https://ndhadeliver.natlib.govt.nz/webarchive/20221122145232/https://digitalcouncil.govt.nz/advice/reports/towards-trustworthy-and-trusted-automated-decision-making-in-aotearoa/

Dijkstra, E. W. (1982). "On the role of scientific thought". In E. W. Dijkstra, *Selected writings on Computing: A Personal Perspective* (pp. 60-66). New York: Springer-Verlag.

Dingli, A., & Farrugia, D. (2023). *Neuro-Symbolic AI.* Packt Publishing.

DPMC. (2025, December 11). *Community engagement* . Retrieved from Department of the Prime Minister and Cabinet: https://www.dpmc.govt.nz/our-programmes/policy-project/policy-tools/community-engagement

DPMC. (2026, February 26). *New Zealand's Cyber Security Strategy 2026 – 2030.* Retrieved from Department of Prime Minister and Cabinet: https://www.dpmc.govt.nz/publications/new-zealands-cyber-security-strategy-2026-2030

Fraser, H. (2021, December 7). *What is Better Rules?* . Retrieved from NZ Digital Government: https://www.digital.govt.nz/blog/what-is-better-rules#:~:text=Simulations:,new%20solutions%20not%20previously%20explored.

Fredrickson, O. (2022). Risk assessment algorithms in the New Zealand criminal justice system. *New Zealand Law Journal*.

Fregly, C., Barth, A., & Eigenbrode, S. (2023, June 20). *Generative AI on AWS.* O'Reilly Media. Retrieved from Flow-AI: https://www.flow-ai.com/blog/improving-llm-systems-with-a-b-testing

Gavaghan, C., Knott, A., Maclaurin, J., Zerelli, J., & Liddicoat, J. (2019). *Government Use of Artificial Intelligence in New Zealand.* University of Otago.

GeeksForGeeks. (2025, July 15). *Understanding Encapsulation, Inheritance, Polymorphism, Abstraction in OOPs* . Retrieved from GeeksForGeeks: https://www.geeksforgeeks.org/java/understanding-encapsulation-inheritance-polymorphism-abstraction-in-oops/

Ghafour, B. (2025, August 20). *A Theory of Information, Variation, and Artificial Intelligence.* Retrieved from https://arxiv.org/html/2508.19264v1

Glassner, A. (2021). *Deep Learning: A Visual Approach.* No Starch Press.

Gluckman, P. (2017, June 19). *Using Evidence To Inform Social Policy: The Role Of Citizen-Based Analytics.* Retrieved from DPMC: https://www.dpmc.govt.nz/sites/default/files/2021-10/pmcsa-17-06-19-Citizen-based-analytics.pdf

Gluckman, P., Ahie, M., Ferguson, M., Hauser, H., Hayden, B., Levin, N., . . . Spencer, H. (2024, August). *Science System Advisory Group Report - An architecture for the future.* Retrieved from MBIE: https://www.mbie.govt.nz/dmsdocument/30024-science-system-advisory-group-report-pdf

Government Chief Digital Officer. (2024, September 16). *Full results: 2024 cross-agency survey of use cases for artificial intelligence (AI)*. Retrieved from digital.govt.nz: https://www.digital.govt.nz/dmsdocument/262~full-results-2024-cross-agency-survey-for-artificial-intelligence-ai-use-cases/html

Gray, H. (2021). Should Māori trust the public health system? *Metro Magazine*. Retrieved from Metro.

Greco, F. (2019). *Traveling Salesman Problem.* Rijeka: INTECH d.o.o.

Green, K. (2024, June 21). *Kāpiti Coast residents demand council throw out report on sea level rise* . Retrieved from Radio New Zealand: https://www.rnz.co.nz/news/national/520144/kapiti-coast-residents-demand-council-throw-out-report-on-sea-level-rise

Health New Zealand | Te Whatu Ora. (2023). *Evaluation of two tools used for waitlist prioritisation for planned care in Health New Zealand – Te Whatu Ora.* Health New Zealand | Te Whatu Ora.

Herries, J. (2025, February 27). *Aide-Memoire - Use of Al-created Clinical Documentation.* Retrieved from Te Whatu Ora: https://www.tewhatuora.govt.nz/assets/Uploads/HNZ00074985-Aide-Memoire-Use-of-AI-driven-Clinical-Documentation.pdf

Hoffmann, A. (2001). Artificial and Natural Computation. In N. J. Smelser, & P. B. Baltes, *International Encyclopedia of the Social & Behavioral Sciences* (pp. 777-783).

Hofmann, V., Kalluri, P. R., Jurafsky, D., & King, S. (2024). AI generates covertly racist decisions about people based on their dialect. *Nature, 633*, 147-154.

Horni, A., Nagel, K., & Axhausen, K. W. (2025, March 24). *The Multi-Agent Transport Simulation.* Retrieved from MATSim.org: https://matsim.org/files/book/partOne-latest.pdf

House of Lords. (2019, March 9). *Regulating in a digital world - 2nd Report of Session 2017–19.* Retrieved from House of Lords: https://publications.parliament.uk/pa/ld201719/ldselect/ldcomuni/299/299.pdf

Ipsos. (2025, August 1). *Ipsos NZ AI Monitor 2025.* Retrieved from Ipsos: https://www.ipsos.com/en-nz/understanding-aotearoa-new-zealand-ipsos-ai-monitor-2025

Jackson, R. (2005, May 30). *Farewell to the Wanganui Computer* . Retrieved from Computerworld: https://www.computerwoche.de/article/2609769/farewell-to-the-wanganui-computer.html

Janssen, M., & Voort, H. v. (2016, January). Adaptive governance: Towards a stable, accountable and responsive government. *Government Information Quarterly, 33*(1), 1-5.

Jenkins, K. (2023, May). Synthetic Data and Public Policy: supporting real-world policymakers with algorithmically generated data. *Policy Quarterly, 19*(2), 29-39.

Jin, C. (2024, August 8). *AI Governance within Health New Zealand.* Retrieved from Victoria University of Wellington School of Engineering and Computer Science: https://ecs.wgtn.ac.nz/foswiki/pub/Groups/AI_and_Society/AI_and_Society_Seminars/AI%20Governance%20within%20Health%20New%20Zealand.pptx

Jolliffe Simpson, A. D., Joshi, C., & Polaschek, D. L. (2021). Predictive Validity of the DYRA and SAFVR: New Zealand Police's Family Violence Risk Assessment Instruments. *Criminal Justice and Behavior, 48*(1), 1487-1508.

Kimi Team. (2025, July 28). *Kimi K2: Open Agentic Intelligence.* Retrieved from arXiv: https://arxiv.org/html/2507.20534v1#S2

Knight, D. (2025). Our constitutional ecosystem: distinctive features and dangerous foes. *Hāpai Public AGM.* Wellington: SSRN.

Lensen, A., McGavin, C., & Mudgway, C. (2025, September 1). *A call to the NZ Parliament to regulate AI* . Retrieved from Regulate AI NZ: https://regulateai.nz

Lucas, R. (1976). Econometric policy evaluation: A critique. *Carnegie-Rochester Conference Series on Public Policy*, (pp. 19-46).

Michalewicz, Z., & Michalewicz, M. (1997). Evolutionary Computation Techniques and Their Applications. *IEEE International Conference on Intelligent Processing Systems* , (pp. 14-25). Beijing.

Microsoft. (2024, May 8). *2024 Work Trend Index Annual Report.* Retrieved from Microsoft: https://assets-c4akfrf5b4d3f4b7.z01.azurefd.net/assets/2024/05/2024_Work_Trend_Index_Annual_Report_663d45200a4ad.pdf

Ministry for Regulation. (2025, July 14). *Official information request - Use of AI for submissions on discussion document.* Retrieved from Ministry for Regulation: https://www.regulation.govt.nz/assets/Publication-Documents/20250714-OIA-Response-Use-of-AI-in-submissions-analysis-policies-and-safeguards-R00979.pdf

Ministry of Justice. (2024, April 11). *Regulatory Impact Statement: Reinstating three strikes sentencing law.* Retrieved from Ministry for Regulation: https://www.regulation.govt.nz/assets/RIS-Documents/ris-justice-rtssl-mar24.pdf#:~:text=Three%20strikes%20is%20a%20sentencing%20model%20originating,through%20the%20guarantee%20of%20tough%20sentencing%20outcomes.

Ministry of Justice. (n.d.). *The Algorithm Charter* . Retrieved from New Zealand Ministry of Justice: https://www.justice.govt.nz/justice-sector-policy/key-initiatives/cross-government/the-algorithm-charter/

Ministry of Social Development. (2021, October). *Model Development Lifecycle* . Retrieved from Ministry of Social Development: https://www.msd.govt.nz/about-msd-and-our-work/work-programmes/initiatives/phrae/model-development-lifecycle.html

Ministry of Transport. (2024, August 20). *Travel demand in New Zealand 2050.* Retrieved from Ministry of Transport Te Manatū Waka: https://consult.transport.govt.nz/policy/long-term-insights-briefing-

consultation/user_uploads/final-ltib-consultation-on-topic-august-2024.pdf#page7

MIT Laboratory for Information and Decision Systems. (2020, October 16). *The real promise of synthetic data.* Retrieved from MIT News: https://news.mit.edu/2020/real-promise-synthetic-data-1016

Murikah, W., Nthenge, J. K., & Musyoka, F. M. (2024). Bias and ethics of AI systems applied in auditing - A systematic review. *Scientific African, 25*, e02281.

National Institute of Standards and Technology. (2024, July). *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile.* Retrieved from NIST Technical Series Publications: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf

Niederhoffer, K., Kellerman, G. R., Lee, A., Liebscher, A., Rapuano, K., & Hancock, J. T. (2025, September 23). AI-Generated "Workslop" Is Destroying Productivity . *Harvard Business Review*.

NZGP. (2019, May). *Syndicated Procurement: Quick Guide.* Retrieved from New Zealand Government Procurement: https://www.procurement.govt.nz/assets/procurement-property/documents/guide-syndicated-procurement.pdf

OECD. (2019, November). Scoping the OECD AI Principles. *OECD Digital Economy Papers*(291). Retrieved from OECD Publishing.

OECD. (2024). *AI principles.* Retrieved from OECD: https://www.oecd.org/en/topics/sub-issues/ai-principles.html

OECD. (2024, March). *Explanatory memorandum on the updated OECD definition of an AI system.* Retrieved from OECD.AI: https://www.oecd.org/content/dam/oecd/en/publications/reports/2024/03/explanatory-memorandum-on-the-updated-oecd-definition-of-an-ai-system_3c815e51/623da898-en.pdf

Ombudsman. (2019, May 8). *Requests for internal decision making rules: A guide to section 22 of the OIA and section 21 of the LGOIMA.* Retrieved from Ombudsman New Zealand: https://www.ombudsman.parliament.nz/resources/requests-internal-decision-making-rules-guide-section-22-oia-and-section-21-lgoima

Ombudsman. (2019, May 6). *Requests for reasons for a decision or recommendation: A guide to section 23 of the OIA and section 22 of the LGOIMA* . Retrieved from Ombudsman New Zealand: https://www.ombudsman.parliament.nz/resources/requests-reasons-decision-or-recommendation-guide-section-23-oia-and-section-22-lgoima

Ombudsman. (2022, September 28). *Ready or not? OIA compliance and practice in 2022* . Retrieved from Ombudsman New Zealand: https://www.ombudsman.parliament.nz/resources/oia-compliance-and-practice-ready-or-not-2022

Ombudsman. (2026, March 9). *The OIA and algorithm-based decision-making*. Retrieved from FYI.org.nz: https://fyi.org.nz/request/33209-ai-systems-and-oia-section-22-and-23

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., . . . Schulman, J. (2022). *Training language models to follow instructions with human feedback.* OpenAI.

Patch, G. (2015). Chapter 11 - Software Design and Development. In G. Patch, K. R. Fowler, & C. L. Silver, *Developing and Managing Embedded Systems and Products* (pp. 502-504). Oxford: Newnes.

PHF Science. (2024, December 9). *The future is now: revolutionising decision-making with AI-driven simulations.* Retrieved from PHF Science: https://www.phfscience.nz/news-publications/the-future-is-now-revolutionising-decision-making-with-ai-driven-simulations/

Privacy Commissioner. (2013). *What's the difference between the Official Information Act and Privacy Act?* Retrieved from Office of the Privacy Commissioner: https://privacy.org.nz/tools/knowledge-base/view/183

Quinlan, J. (1986). Induction of Decision Trees. *Machine Learning, 1*, 81-106.

Russell, S., Perset, K., & Grobelnik, M. (2023, November 29). *Updates to the OECD's definition of an AI system explained.* Retrieved from OECD.AI: https://oecd.ai/en/wonk/ai-system-definition-update

Schwarcz, D., & Prince, A. E. (2020). Proxy Discrimination in the Age of Artificial Intelligence and Big Data. *Iowa Law Review, 105*, 1257.

Sepuloni, C. (2018, May 10). *New Privacy, Human Rights and Ethics framework essential step in safe data use*. Retrieved from Beehive: https://www.beehive.govt.nz/release/new-privacy-human-rights-and-ethics-framework-essential-step-safe-data-use

Shaw, J. (2020, July 28). *New Algorithm Charter a world-first* . Retrieved from Beehive.govt.nz: https://www.beehive.govt.nz/release/new-algorithm-charter-world-first

Simpson Grierson. (2023, May 19). *A class of their own? Class actions, privacy breaches and what it could mean for you* . Retrieved from Simpson Grierson: https://www.simpsongrierson.com/insights-news/legal-updates/a-class-of-their-own-class-actions-privacy-breaches-and-what-it-could-mean-for-you

Social Investment Agency. (2018, December). *From listening to learning.* Retrieved from Social Investment Agency: https://www.sia.govt.nz/assets/Uploads/From-Listening-to-Learning.pdf

Social Investment Unit. (2017). *Measuring the fiscal impact of social housing services – Technical report.* Wellington, New Zealand.

Social Wellbeing Agency. (2021, December). *Data Protection and Use Policy (DPUP).* Retrieved from digital.govt.nz: https://www.digital.govt.nz/standards-and-guidance/privacy-security-and-risk/privacy/data-protection-and-use-policy-dpup

Sohl-Dickstein, J. (2022, November 6). *Too much efficiency makes everything worse: overfitting and the strong version of Goodhart's law* . Retrieved from Jascha's blog: https://sohl-dickstein.github.io/2022/11/06/strong-Goodhart.html#endnote-overfittinggenerality

StatsNZ. (2018, October). *Algorithm assessment report 2018.* Retrieved from data.govt.nz: https://data.govt.nz/toolkit/data-ethics/government-algorithm-transparency-and-accountability/algorithm-assessment-report/

StatsNZ. (2020, July). *Algorithm charter for Aotearoa New Zealand* . Retrieved from data.govt.nz: https://data.govt.nz/toolkit/data-ethics/government-algorithm-transparency-and-accountability/algorithm-charter

StatsNZ. (2020, November 23). *Ngā Tikanga Paihere.* Retrieved from data.govt.nz: https://data.govt.nz/toolkit/data-ethics/nga-tikanga-paihere

StatsNZ. (2021, September). *The system strategy - The Government Data Strategy and Roadmap*. Retrieved from data.govt.nz: https://data.govt.nz/leadership/strategy-and-roadmap/strategy

StatsNZ. (2022, August). *About data in the IDI - Integrated Data Infrastructure*. Retrieved from StatsNZ: https://www.stats.govt.nz/integrated-data/integrated-data-infrastructure/#about

StatsNZ. (2022, August). *How we keep integrated data safe* . Retrieved from StatsNZ: https://www.stats.govt.nz/integrated-data/how-we-keep-integrated-data-safe/

StatsNZ. (2024, May 29). *Linking 2023 Census responses to the Integrated Data Infrastructure.* Retrieved from StatsNZ: https://www.stats.govt.nz/methods/linking-2023-census-responses-to-the-integrated-data-infrastructure/

StatsNZ. (2025, November 19). *Privacy statement and submissions analysis using AI* . Retrieved from StatsNZ: https://www.stats.govt.nz/consultations/public-consultation-proposed-data-collection-approach-and-content-for-the-census/privacy-statement-and-submissions-analysis/

StatsNZ and OPC. (2018, May). *Principles for the safe and effective use of data and analytics.* Retrieved from StatsNZ: https://www.stats.govt.nz/assets/Uploads/Data-leadership-fact-sheets/Principles-safe-and-effective-data-and-analytics-May-2018.pdf

Stevenson, J. (2019, July). *Exploring Machine Consumable Accident Compensation Legislation.* Retrieved from Service Innovation Lab: https://serviceinnovationlab.github.io/assets/Exploring_Machine_Consumable_Code_With_ACC.pdf

Stewart, B. (1998). *Necessary and Desirable – Privacy Act 1993 Review.* Retrieved from https://www.abuseincare.org.nz/__data/assets/pdf_file/0025/27592/stewart-b-necessary-and-desirable-privacy-act-1993-review-office-of-the-privacy-commissioner-1997.pdf

Stigler, S. M. (1981). Gauss and the Invention of Least Squares. *The Annals of Statistics, 9*(3), 465-474.

Stix, C. (2021). Actionable Principles for Artificial Intelligence Policy: Three Pathways. *Science and Engineering Ethics, 27*(1), 15.

Taylor Fry. (2021, December 20). *Algorithm Charter for Aotearoa New Zealand Year 1 Review.* Retrieved from data.govt.nz: https://data.govt.nz/assets/data-ethics/algorithm/Algorithm-Charter-Year-1-Review-FINAL.pdf

Taylor Fry. (2021, June). *NZ Police - Safe and ethical use of algorithms.* Retrieved from NZ Police: https://www.police.govt.nz/sites/default/files/publications/safe-ethical-use-algorithms-report.pdf

Te Kāhui Raraunga. (2023, May). *Māori Data Governance Model.* Retrieved from kahuiraraunga.io: https://www.kahuiraraunga.io/_files/ugd/b8e45c_a5b7af8b688c4cd9b7583775c27da52e.pdf

Te Kāhui Raraunga. (2025). *Māori Artificial Intelligence Governance Framework. Contextualised advice for AI use, extending the Māori data governance model.* Retrieved from Te Kāhui Raraunga: https://www.kahuiraraunga.io/maoriaigovernance

Te Mana Raraunga. (2018, October). *Principles of Māori Data Sovereignty.* Retrieved from Te Mana Raraunga: https://static1.squarespace.com/static/58e9b10f9de4bb8d1fb5ebbc/t/5bda208b4ae237cd89ee16e9/1541021836126/TMR+Māori+Data+Sovereignty+Principles+Oct+2018.pdf

The Generator. (2024, October 20). *NZ's GovGPT chatbot is a perfect example of how to do trustworthy AI.* Retrieved from Medium: https://medium.com/the-generator/how-govgpt-nz-sets-ai-safeguard-standards-a666d3ac170b

Tom, E., Aurum, A., & Vidgen, R. (2013, June). An exploration of technical debt. *Journal of Systems and Software, 86*(6), 1498-1516.

Treasury. (2018, May 3). *Treasury Report: Upcoming Cabinet Paper: Fair Pay Agreements.* Retrieved from Treasury: https://www.treasury.govt.nz/sites/default/files/2018-09/oia-20180336.pdf

Treasury. (2024, August). *Tax and Welfare Analysis (TAWA) Model Methodology Report.* Retrieved from The Treasury New Zealand: https://www.treasury.govt.nz/publications/guide/tax-and-welfare-analysis-tawa-model-methodology-report

Treasury Board of Canada Secretariat. (2023, April 25). *Directive on Automated Decision-Making* . Retrieved from Canada.ca: https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32592

Tucci, C., Della Greca, A., Tortora, G., & Francese, R. (2024). Explainable biometrics: a systematic literature review . *Journal of Ambient Intelligence and Humanized Computing*.

Turpin, M., Michael, J., Perez, E., & Bowman, S. R. (2023). Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. *Proceedings of the 37th International Conference on Neural Information Processing Systems* (pp. 74952 - 74965). New Orleans: Curran Associates Inc.

Tweedie, F. (2023, December). *Algorithm Impact Assessment toolkit* . Retrieved from data.govt.nz: https://data.govt.nz/toolkit/data-ethics/government-algorithm-transparency-and-accountability/algorithm-impact-assessment-toolkit

Verian. (2026, March 2). *Aotearoa Internet Insights 2025.* Retrieved from InternetNZ: https://internetnz.nz/new-zealands-internet-insights/new-zealands-internet-insights-2025/

Vowles, P. (2021, May 5). *Automated Decision Making Standard* . Retrieved from Ministry of Social Development: https://www.msd.govt.nz/documents/about-msd-and-our-work/publications-resources/official-information-responses/2022/july/07072022-requesting-the-document-automated-decision-making-in-msd-proposed-legislative-and-policy-framework-memo-.pdf

Waitangi Tribunal. (2005). *The Offender Assessment Policies Report Wai 1024.* . Wellington, New Zealand: Legislation Direct.

Waitangi Tribunal. (2021). *The Report on the Comprehensive and Progressive Agreement for Trans-Pacific Partnership.* Retrieved from Ministry of Justice: https://forms.justice.govt.nz/search/Documents/WT/wt_DOC_195473606/Report%20on%20the%20CPTPP%20W.pdf

Walters, L. (2025, February 21). *Stanford writes to public service chief over botched teacher supply data* . Retrieved from Newsroom: https://newsroom.co.nz/2025/02/21/stanford-writes-to-public-service-chief-over-botched-teacher-supply-data/

Whittaker, R., Dobson, R., Jin, C. K., Style, R., Jayathissa, P., Hiini, K., . . . Muir, P. (2023). An example of governance for AI in health services from Aotearoa New Zealand. *npj Digital Medicine, 6*(164).

Willis, N. (2024, May). *Accelerating Social Investment.* Retrieved from Social Investment Agency: https://www.sia.govt.nz/social-investment

Wilson, D., Tweedie, F., Rumball-Smith, J., Ross, K., Kazemi, A., Galvin, V., & Blakey, J. (2022). Lessons learned from developing a COVID-19 algorithm governance framework in Aotearoa New Zealand. *Journal of the Royal Society of New Zealand, 53*(1), 82-94.

Winder, P. (2020). *Reinforcement Learning.* O'Reilly Media.

Woods, A. (2024, May 2). *OIA-2024-4974.* Retrieved from New Zealand Defence Force: https://www.nzdf.mil.nz/assets/Uploads/DocumentLibrary/OIA-2024-4974-Charter-algorithms.pdf

Yao, Z., Liu, Y., Chen, Y., Chen, J., Fang, J., Hou, L., . . . Chua, T.-S. (2025). *Are Reasoning Models More Prone to Hallucination?* Retrieved from https://arxiv.org/abs/2505.23646

Yarwood, V. (2014). A Bomb for 'Big Brother'. *New Zealand Geographic*, 125-.

Zerilli, J., Knott, A., Maclaurin, J., & Gavaghan, C. (2019). Algorithmic Decision-Making and the Control Problem. *Minds and Machines, 29*, 555-578.

Zheng, S., Trott, A., Srinivasa, S., Parkes, D. C., & Socher, R. (2022). The AI Economist: Taxation policy design via two-level deep multiagent reinforcement learning. *Science Advances*.