

DISCUSSION PAPER:

Applying OECD AI Principles and Recommendations in Aotearoa New Zealand

Implications from a Research Perspective

AUGUST 2026



ABOUT

The present paper was prepared by Dr Alvaro Orsi (formerly of PHF Science; currently Enterprise AI Architect at Contact Energy), Professor Albert Bifet (University of Waikato), Professor Gillian Dobbie (University of Auckland), Professor Stephen Cranefield (University of Otago).

The views and opinions expressed in this discussion paper are those of the authors listed above and, where referenced, participants in the panel discussions described herein. They do not reflect the views of any organisation or person, whether affiliated with the authors or participants or otherwise.

CIRCULATED BY

The Artificial Intelligence Researchers
Association Incorporated
Website: www.airesearchers.nz
Email: hello@airesearchers.nz

CONTENTS

EXECUTIVE SUMMARY	4
INTRODUCTION	5
OECD AI Principles Overview	6
Key Recommendations from National AI Strategic Documents	7-8
APPLICATION OF OECD AI PRINCIPLES IN AOTEAROA NEW ZEALAND	9
1: Inclusive Growth, Sustainable Development & Wellbeing	10-12
2: Respect for the Rule of Law, Human Rights & Democratic Values, including Fairness & Privacy (Human-Centered Values & Fairness)	13-15
3: Transparency & Explainability	16-18
4: Robustness, Security & Safety	19-21
5: Accountability	22-24
CONCLUSION	25
The Role of AI Researchers for implementing OECD AI Principles in Aotearoa	25
AI Research Nexus Panel Recommendations	26-27
Conference Panel Discussion: Implementing Strategy & Principles in Aotearoa	28
REFERENCES	29

Executive Summary

The OECD AI Principles provide a critical framework for AI researchers in Aotearoa New Zealand, serving dual functions of empowerment and obligation. These principles – encompassing inclusive growth, human-centred values, transparency, robustness, and accountability – enable researchers to advocate for ethical AI development whilst establishing clear research responsibilities.

From a research perspective, the principles both support and constrain AI development activities. They empower researchers to pursue work that prioritises societal benefit, environmental sustainability, and cultural values, including Te Tiriti o Waitangi and Māori data sovereignty. This legitimises research focused on addressing inequalities, climate change, and social well-being through AI technologies.

Simultaneously, the principles establish research obligations requiring integration of fairness assessments, explainability mechanisms, security protocols, and accountability frameworks into research methodologies. Researchers must navigate bias mitigation, develop transparent AI systems, ensure robust security measures, and establish clear accountability mechanisms.

For Aotearoa New Zealand's AI research landscape, this creates opportunities to lead in developing culturally responsive AI frameworks that demonstrate how global standards can be operationalised within indigenous and bicultural contexts. The principles align with national initiatives including the Government's Approach to Work on Artificial Intelligence [1, 2], the AI Researchers Association's 2024 whitepaper Aotearoa NZ

AI: A Strategic Approach [3], and international commitments such as the Bletchley Declaration.

This positioning enables Aotearoa New Zealand researchers to contribute meaningfully to international AI standards whilst ensuring development reflects the nation's unique values and commitments.

This document also draws on insights from two panel sessions which took place in 2025:

- *"From Principles to Practice: OECD AI Principles as a Strategic Framework for AI Research"*, held on 26 June at the AI Research Nexus Conference.
- *"Translating OECD AI Principles into research action: challenges and opportunities for AI researchers"*, held on 18 November at the AI Researchers Association Annual Conference.

These discussions reinforced the need to engineer trust (e.g., applicability domains, adversarial robustness, and bias evaluation), treat AI as civic infrastructure with public-interest oversight, embed Māori data sovereignty through co-design and epistemic humility, and prioritise smaller, domain-specific, interpretable systems deployed on sovereign infrastructure for sensitive data.

Introduction

Artificial intelligence is reshaping industries, governments, and societies across the globe. Countries like Aotearoa New Zealand, with its strong values-based governance and commitment to inclusivity, face both opportunities and challenges in the responsible use of AI. **The OECD AI Principles [4], adopted in 2019 and updated in 2024, provide a globally recognised ethical framework for nations and organisations to follow.** These principles emphasise that AI should benefit people and the planet, and are designed to support trustworthy and ethical AI development and deployment.

In Aotearoa New Zealand, AI is seen as a powerful tool to drive innovation, address societal inequalities, and tackle pressing issues such as climate change. However, it is essential to approach AI development in a manner that is consistent with the country's values, particularly respecting Māori data sovereignty, human rights, and Te Tiriti o Waitangi. The Ministry of Business, Innovation & Employment (MBIE) has agreed in its "Approach to work on Artificial Intelligence" cabinet paper [1] to promote the OECD AI Principles:

We should promote the OECD AI Principles as a key direction for our approach to responsible AI in New Zealand. The OECD AI Principles were the first intergovernmental agreement on AI and continue to be influential. They represent common ground among New Zealand's trusted partners. They can serve as a visible signal of our national approach and a high-level direction for work across government on AI issues, including in the public service.

This report delves into each of the OECD AI Principles and evaluates their application in Aotearoa New Zealand, while also outlining how research efforts can support and enhance these principles.

OECD AI Principles Overview

The OECD AI Principles [4] are divided into two main categories: value-based principles for responsible AI and recommendations for national policies and international cooperation.

THE OECD RECOMMENDATIONS FOR NATIONAL POLICIES AND INTERNATIONAL COOPERATION INCLUDE:



Investing in AI research and development (R&D):

Promoting AI research to foster innovation while ensuring that ethical considerations are integrated into development practices.



Fostering an inclusive, AI-enabling ecosystem:

Encouraging the development of technical infrastructure, talent, and resources to support AI advancement.



Shaping an enabling interoperable governance and policy environment for AI:

Creating flexible and interoperable governance frameworks that foster trustworthy AI innovation whilst protecting human rights, privacy, safety, security, and wider societal interests.



Building human capacity and preparing for labour market transformation:

Equipping the workforce with the necessary skills to interact with AI technologies and mitigate potential disruptions in the job market.



International co-operation for trustworthy AI:

Engaging in cross-border collaborations and adhering to international standards to ensure that AI development is globally aligned and ethically sound.

THE FIVE CORE OECD AI PRINCIPLES FOR RESPONSIBLE AI DEVELOPMENT ARE:

1

Inclusive Growth, Sustainable Development, and Wellbeing

AI should drive economic growth, while promoting human wellbeing and addressing inequalities. It should be developed and deployed in a way that enhances the wellbeing of individuals and communities, and support sustainable development goals.

2

Respect for the Rule of Law, Human Rights and Democratic Values, including Fairness and Privacy

AI should be developed consistent with human-centred values, such as human rights and democratic values, including freedom, dignity, and privacy. Fairness should be a key component of AI decision-making, ensuring that AI systems do not reinforce biases or discriminatory outcomes.

3

Transparency and Explainability

The processes behind AI decision-making should be transparent and understandable to users, allowing stakeholders to comprehend and trust AI outputs.

4

Robustness, Security, and Safety

AI systems should be secure, resilient, and safe, with protections against cyberattacks, system failures, and unintended consequences.

5

Accountability

Clear mechanisms of accountability should be in place for AI developers and users, ensuring that responsibilities for AI-related outcomes are properly assigned.

Key Recommendations

FROM NATIONAL AI STRATEGIC DOCUMENTS

Three pivotal documents provide comprehensive frameworks for advancing AI capabilities in Aotearoa New Zealand: the AI Researchers Association's 2024 whitepaper Aotearoa NZ AI: A Strategic Approach [3], the AI Forum's blueprint [5], and MBIE's national

AI strategy [6]. These complementary frameworks outline pathways for positioning Aotearoa New Zealand as a responsible global leader in artificial intelligence.



MBIE NATIONAL AI STRATEGY: ADOPTION-FOCUSED APPROACH

MBIE published Aotearoa New Zealand's first national AI strategy in July 2025, titled "Investing with Confidence". This strategy deliberately emphasises AI adoption and application rather than foundational AI development, reflecting the country's economic realities and strategic positioning as a small, advanced economy. Aotearoa New Zealand's competitive advantage lies in becoming sophisticated adopters who can rapidly identify, adapt, and deploy AI solutions to local challenges across agriculture, healthcare, education, and other strategic sectors.

The strategy addresses five interconnected barriers preventing broader AI adoption: regulatory uncertainty, perceived complexity and ethical concerns, lack of understanding and perceived value, skills shortages, and inadequate AI capabilities. To overcome these obstacles, the Government commits to maintaining stable, light-touch, principles-based policy

approaches, publishing responsible AI guidance for businesses, building domestic workforce capacity through university programmes and training initiatives, and leading by example through responsible AI use and adoption across government.

The strategy positions international collaboration as essential for enhancing Aotearoa New Zealand's AI capabilities whilst attracting investment and expertise. The Government also announced significant reforms to the Science, Innovation and Technology system, including creating a new advanced technology public research organisation and introducing a national policy for intellectual property management in universities and public research organisations to help bring AI technologies to market. These measures demonstrate alignment with OECD recommendations for investing in AI research and development, fostering inclusive AI-enabling ecosystems, building human capacity, and promoting international cooperation for trustworthy AI.



AI RESEARCHERS ASSOCIATION STRATEGIC RECOMMENDATIONS

The AI Researchers Association's 2024 whitepaper Aotearoa NZ AI: A Strategic Approach [3] establishes seven core recommendations to strengthen Aotearoa New Zealand's AI research ecosystem. The foundation centres on substantial investment in AI research infrastructure, including creating networks of research centres and doubling the number of AI researchers by 2033, with particular emphasis on promoting Māori-led AI research initiatives.

Talent development forms another critical pillar, encompassing comprehensive AI education improvements from primary through postgraduate levels whilst attracting top international talent to strengthen the domestic research base.

The approach advocates for supporting industrial AI adoption through targeted programmes that boost AI implementation in strategic sectors and foster innovative start-ups.

Ethical frameworks receive particular attention, with recommendations to establish culturally appropriate standards for ethical AI use that respect Aotearoa New Zealand's unique cultural context. Data infrastructure development emphasises building national capabilities that support Māori data sovereignty and provide AI testing environments. The strategic approach also prioritises leading AI implementation in government to enhance public sector efficiency whilst ensuring compliance with Te Tiriti o Waitangi. Finally, it emphasises promoting inclusion and social wellbeing by empowering diverse communities, particularly Māori, to engage meaningfully with AI technologies.



AI FORUM BLUEPRINT OBJECTIVES

The AI Forum's Blueprint for Aotearoa [5] presents six strategic objectives that complement the research-focused recommendations. Building trust and knowledge constitutes the primary objective, fostering public confidence through promoting safe and ethical AI use whilst ensuring transparent and accountable adoption practices.

The blueprint emphasises encouraging AI adoption across industries of all sizes, facilitating integration that enhances productivity and drives innovation. Central to this approach is upholding Te Tiriti o Waitangi by ensuring Māori voices are incorporated into AI governance, particularly regarding algorithm sovereignty and decision-making processes.

Workforce development receives significant attention through upskilling initiatives designed to create a diverse and inclusive workforce capable of leveraging AI technologies effectively. Environmental considerations are addressed through promoting sustainable AI development that supports environmentally sound technologies and infrastructure aligned with national environmental goals. The final objective focuses on strengthening research and investment by encouraging robust research pipelines supported by active investors to sustain AI innovation and implementation.

Collectively, these frameworks provide comprehensive guidance for advancing AI capabilities whilst maintaining ethical standards and cultural appropriateness essential for responsible AI development in Aotearoa New Zealand.

Application of OECD AI Principles in Aotearoa New Zealand

This section examines how each of the OECD AI Principles can be applied within the unique social, cultural, and economic context of Aotearoa New Zealand. We explore how existing initiatives align with these principles and identify areas where further research and actions are needed.

1

INCLUSIVE GROWTH, SUSTAINABLE DEVELOPMENT & WELLBEING

The OECD's first principle mandates that AI should drive economic growth whilst promoting human wellbeing and addressing inequalities. This principle requires AI to be developed and deployed in ways that enhance the wellbeing of individuals and communities whilst supporting sustainable development goals. In Aotearoa New Zealand's context, this principle emphasises leveraging AI to promote inclusive economic growth and address societal challenges, with particular attention to environmental sustainability and respect for Māori data and algorithmic sovereignty.

Aotearoa New Zealand can operationalise inclusive growth by prioritising mission-led, domain-specific AI that delivers small, compounding wins in areas of national advantage – primary industries, climate resilience, health, and te reo Māori – rather than chasing frontier-scale models. Compact, interpretable systems that combine classical methods (e.g., computer vision) with modern generative techniques can solve real problems efficiently, reduce compute and carbon costs, and build public trust through clarity and performance

proportionality. Treating AI as civic infrastructure – co-governed, research-accessible, and aligned with public-good mandates – helps avoid the pitfalls of closed, commercial platforms and anchors benefits in communities, workforces, and services.

Cultural legitimacy is central to wellbeing. Generic generative outputs risk sanitised or colonised representations; co-design, Māori-led research, and data and algorithmic sovereignty safeguards are essential to ensure nuance, authority, and benefit sharing. Pragmatic pathways include open-source models and tooling, domestic infrastructure for sensitive data, and a focus on small-data, high-impact applications that improve local processes and services. Finally, embedding governance for evidence traceability and domain-grounded fact-checking equips organisations to counter confident misuse of AI and to channel productivity gains into equitable outcomes. Taken together, these approaches align environmental sustainability, social equity, and economic productivity – maximising inclusive growth in ways that reflect Aotearoa's values, constraints, and strengths.



PATHWAY TO IMPLEMENTATION

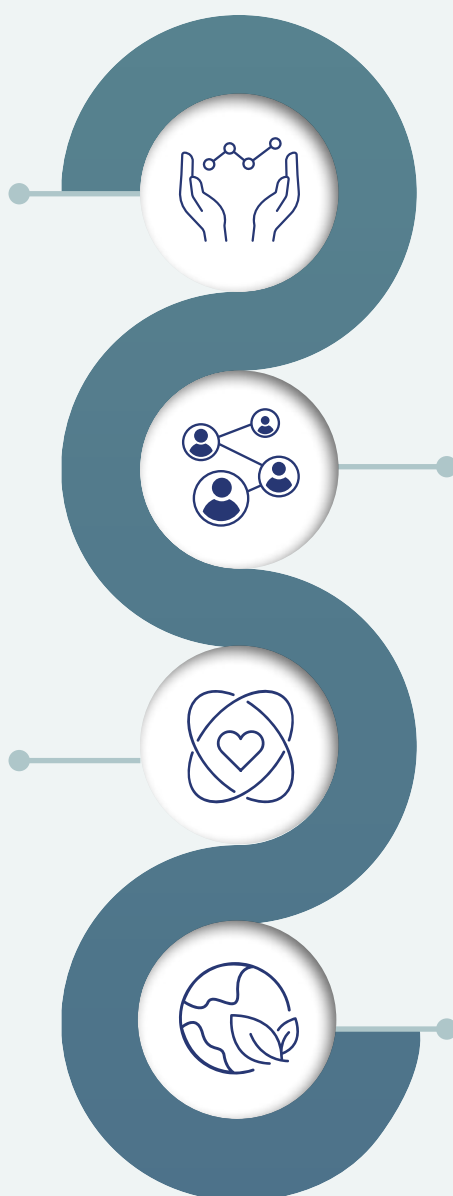
Implementation of this principle in Aotearoa New Zealand requires a multi-faceted research approach centred on four core pathways.

MĀORI DATA SOVEREIGNTY

Respecting Māori data sovereignty is paramount through initiatives like Te Mana Raraunga [7], which advocate for Māori control over data pertaining to them, ensuring AI applications do not infringe upon their rights [8].

AI FOR SOCIAL AND ENVIRONMENTAL SUSTAINABILITY

Researchers should embrace opportunities to work collaboratively with local and national communities to apply AI to help address social and environmental problems, especially those aligned with the United Nations' Sustainable Development Goals (SDGs) [9] and in line with best practice [10].



ACCESSIBILITY OF AI

Researchers should seek to increase the accessibility of AI in Aotearoa New Zealand by engaging in outreach with diverse communities.

SUSTAINABLE DEVELOPMENT OF AI¹

AI research in Aotearoa New Zealand should pay attention to the energy and environmental costs of the specific technologies and services used, whether running in Aotearoa New Zealand or in an overseas data centre. Researchers should consider the trade-offs in performance vs. environmental costs that they are making, and consider tackling research problems in green AI implementation, monitoring and governance.

¹ While OECD AI Principle 1 focuses solely on applying AI to enhance sustainability and wellbeing in society, it is also crucial to ensure that AI development and deployment are carried out sustainably. This need is recognised in clause 2.4 of the Seoul Ministerial Statement for advancing AI safety, innovation and inclusivity, 2024 [11], which New Zealand is a signatory of.

RESEARCH IMPLICATIONS

This principle establishes several critical research imperatives for Aotearoa New Zealand's AI community:

- **Zero-shot learning in research** to address sustainability and reusability challenges
- **Continual learning and fine-tuning models** for ongoing improvement
- **Developing AI models that integrate indigenous knowledge and perspectives** to ensure cultural appropriateness
- **Developing technologies to assess the societal impact of AI** to prevent exacerbating inequalities
- **Promoting research that addresses organisational intelligence** to create efficiencies in organisations
- **Promoting research that addresses environmental challenges through AI** as a national priority
- **Making fundamental AI research developed in Aotearoa New Zealand open source** to contribute to the democratisation of algorithms and technology
- **Focusing on Green AI efforts** to reduce the carbon impact of AI models when deployed at scale

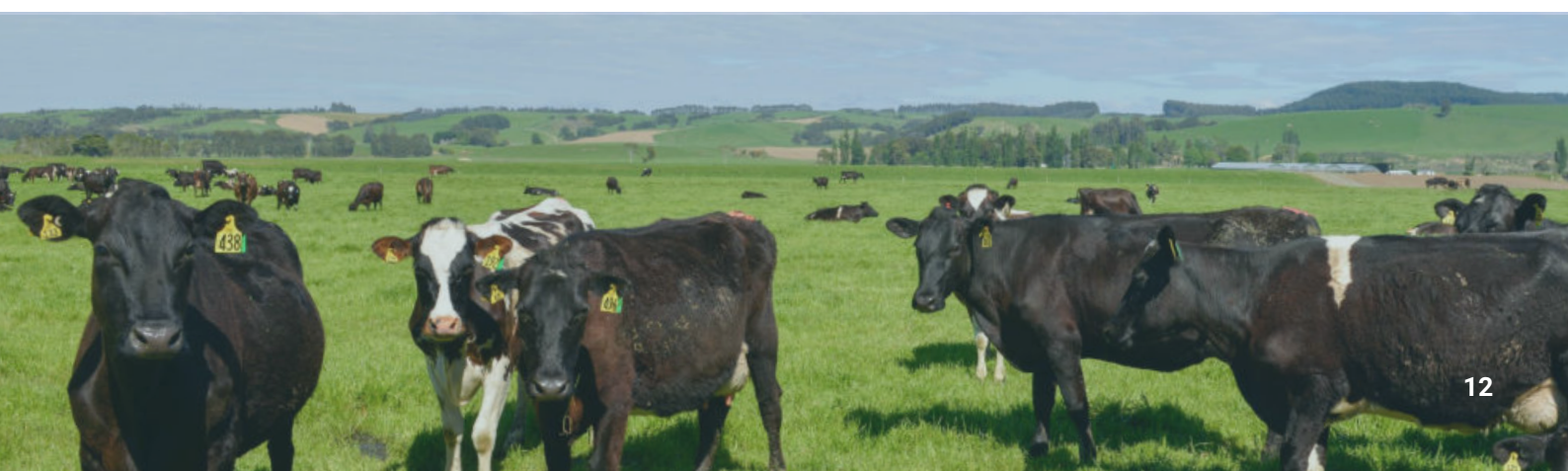
CASE STUDY: UNIVERSITY OF OTAGO METHANE EMISSIONS MONITORING

A PhD project conducted jointly by the University of Otago's Schools of Computing and Surveying exemplifies this principle's implementation [12]. The research developed a neural network approach to measuring methane emissions from Aotearoa New Zealand agriculture using satellite data. Given that methane is a powerful greenhouse gas, accurate measurement represents a necessary step towards targeted emission reductions.

The project addressed a significant technical challenge: whilst current satellite-based technologies can determine the point source of methane from large outputs such as industrial

leaks, weaker and widely distributed sources from agriculture require a computationally expensive inverse model to approximate area source locations. The project addressed this issue by simulating inverse models directly using a neural network, allowing rapid area assessment of methane emissions.

This work received recognition by winning the SpaceBase Space for Planet Earth Challenge 2023-24, with judges praising the methodology, the prototype system for interactive spatial querying, and its potential applicability to next-generation satellites. This case study demonstrates how AI research can simultaneously address environmental sustainability goals whilst contributing to Aotearoa New Zealand's agricultural sector and international climate commitments.



2

RESPECT FOR THE RULE OF LAW, HUMAN RIGHTS AND DEMOCRATIC VALUES, INCLUDING FAIRNESS AND PRIVACY

The OECD's second principle mandates that AI should be developed consistent with human-centred values, including human rights and democratic values such as freedom, dignity, and privacy. This principle requires that fairness be a key component of AI decision-making, ensuring that AI systems do not reinforce biases or produce discriminatory outcomes. In Aotearoa New Zealand's context, this principle emphasises ensuring AI systems uphold human rights and democratic values as a national priority, with particular attention to Te Tiriti o Waitangi obligations and the protection of vulnerable communities from algorithmic discrimination.

Advancing human-centred AI in Aotearoa requires shifting from generic "responsible AI" rhetoric to concrete engineering and governance practices tailored to local contexts. Fit-for-purpose transparency is essential: disclosures should be designed for the intended audience — users, regulators, domain experts — so they can understand limitations, trace evidence, and challenge outcomes without being overwhelmed by cosmetic visuals that obscure detail. Fairness work should prioritise New Zealand-specific harms and demographics rather than importing overseas benchmarks; that

includes building local evaluation datasets, adopting interpretable metrics, and supporting multi-group assessments beyond binary categories, with clear procedures for mitigation and ongoing monitoring.

Cultural integrity is central to human rights and dignity. General-purpose generative systems often produce sanitised or colonised outputs when engaging with te ao Māori and te reo Māori, undermining Māori data sovereignty. Co-design, Māori-led research, and governance that embeds tikanga and Te Tiriti principles are necessary for legitimacy, nuance, and equitable benefit sharing. Privacy protections should be operationalised through sovereign data pathways and on-shore processing for sensitive contexts, coupled with rigorous access controls and audit trails.

Finally, introduce lightweight, risk-based guardrails for AI use in professional settings: evidence traceability, domain-grounded fact-checking, and peer review of AI-assisted decisions. This reduces confident misuse, preserves individual autonomy in AI-supported workflows, and aligns innovation with democratic values, dignity, and justice.

PATHWAY TO IMPLEMENTATION

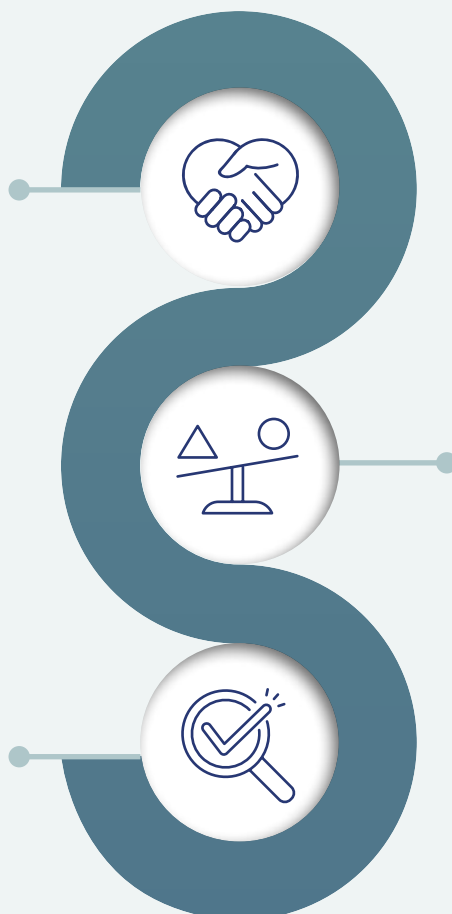
Implementation of this principle requires a comprehensive research approach centred on three core pathways.

ETHICAL FRAMEWORKS

The establishment of culturally appropriate ethical guidelines ensures AI respects human dignity and cultural values, incorporating Te Tiriti o Waitangi principles and Māori worldviews into AI development processes.

RESPONSIBLE RESEARCH AND INNOVATION (RRI)

This approach balances commercial and other goals with those concerned with wider wellbeing, ensuring research considers societal implications throughout the development process.



BIAS MITIGATION

Systematic efforts are underway to prevent AI systems from perpetuating biases, particularly those affecting marginalised communities, through the development of bias detection and mitigation techniques.

RESEARCH IMPLICATIONS

This principle establishes several critical research imperatives for Aotearoa New Zealand's AI community:

- **Investigating the role of automated decision support systems** and how to preserve individual autonomy within AI-assisted decision-making processes
- **Investigating methods to detect and mitigate biases in AI algorithms** through comprehensive AI fairness research
- **Collaborating with ethicists, legal experts, and community leaders** to develop fair AI systems that respect diverse perspectives and values
- **Understanding the impact of AI on different societal groups** to inform equitable policies and prevent discriminatory outcomes
- **Investigating collaborative AI fairness assessment** as part of standard practice when deploying models that have social impact
- **Developing frameworks for preserving privacy** whilst enabling beneficial AI applications
- **Ensuring compliance with Aotearoa New Zealand's privacy legislation** and data protection principles in AI system design

CASE STUDY:

A COLLABORATIVE STUDY DETECTING BIAS IN LLMs

The social inequities experienced by under-represented indigenous populations, such as Māori in Aotearoa New Zealand, are significant compared to non-Indigenous populations. Large Language Models (LLMs) inherit such biases, reflecting the differing treatment of social groups from historical and structural power asymmetries.

Research conducted at the University of Auckland and the University of Waikato in collaboration with students and international researchers addressed identifying, measuring and mitigating bias in LLMs [13]. The project developed an approach for identifying bias in LLMs in the Aotearoa New Zealand context. The involvement of students in the project builds specialised knowledge and capability in this important area.

The project addressed significant technical and social challenges. International solutions do

not work in the Aotearoa New Zealand context. They typically concentrate on demographics of countries like the US, use a diverse range of metrics for evaluation, and are limited to considering bias between two groups, e.g., male and female. To overcome the challenges, the researchers developed a benchmark dataset for training for the Aotearoa New Zealand context, they defined a standard interpretable evaluation metric, and built a tool that can generate synthetic test datasets and compare the bias of different LLMs. The tool is customisable and considers bias across multiple groups, e.g., a range of genders.

This project represents an important step toward ensuring fairness and accountability in LLMs used within Aotearoa New Zealand. As LLMs become more prevalent in decision-making processes, the ability to identify and mitigate bias will be critical to protecting under-represented communities and ensuring equitable outcomes across Aotearoa New Zealand.



3

TRANSPARENCY & EXPLAINABILITY

The OECD's transparency and explainability principle mandates that AI actors should commit to transparency and responsible disclosure regarding AI systems. This principle requires that stakeholders can understand and explain how AI systems make decisions, with clear insights into underlying algorithms, data used, and reasoning behind outputs. In Aotearoa New Zealand's context, transparency involves ensuring public sector AI applications are developed with clarity, allowing citizens to understand and trust AI-driven services whilst maintaining accountability to democratic values and Te Tiriti o Waitangi obligations.

Under this principle, it helps to treat transparency as contextual and audience-specific rather than uniform. Different stakeholders need different forms of clarity: operators may need diagnostics and failure conditions; regulators may need legal rationale and audit trails; the public may need plain-language summaries that convey limits and recourse. Avoid decorative explanations that look

compelling but obscure detail; usefulness hinges on how explanations will be used in practice.

Consider the provenance and evidential basis of outputs. Surfacing data sources, applicable domain standards, and uncertainty signals supports informed judgment and curbs confident misuse of model outputs. Framing transparency as part of a wider assurance process — fact-checking, peer validation pathways, and clear handoffs to human oversight — helps preserve autonomy in decision contexts.

In Aotearoa, transparency also sits within tikanga and Te Tiriti commitments. Co-design with Māori helps prevent generic, sanitised representations that strip nuance, and supports culturally legitimate disclosure practices. Where feasible, smaller, interpretable or hybrid (symbolic plus statistical) approaches can make reasoning pathways more auditable without sacrificing contextual relevance.



PATHWAY TO IMPLEMENTATION

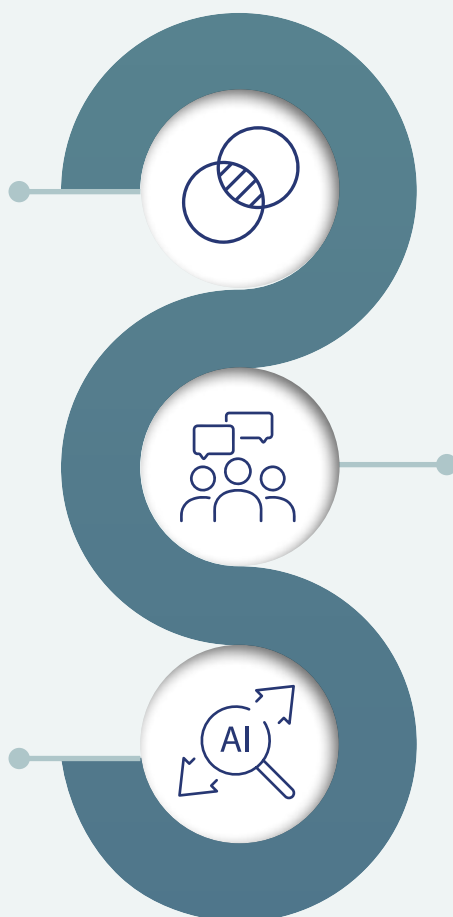
Implementation of this principle requires a comprehensive research approach centred on three core pathways.

GOVERNMENT TRANSPARENCY

Public sector AI applications must be developed with transparency mechanisms that allow citizens to understand and trust AI-driven services, ensuring compliance with democratic accountability standards.

EXPLAINABLE AI RESEARCH

Development of methodologies that make complex AI model outputs understandable to diverse stakeholders, ensuring AI systems can provide clear reasoning for their decisions.



PUBLIC ENGAGEMENT

Educational initiatives are essential to increase public understanding of AI technologies and their implications, fostering informed participation in AI governance decisions.

RESEARCH IMPLICATIONS

This principle establishes several critical research imperatives for Aotearoa New Zealand's AI community:

- **Advancing explainable AI to make complex model outputs understandable** across diverse user groups and technical capabilities
- **Developing tools for auditing AI systems for transparency and accountability** to ensure ongoing compliance with ethical standards
- **Engaging in interdisciplinary research to communicate AI developments effectively** across technical and non-technical audiences
- **Leading implementation of explainable AI model outputs** as New Zealand researchers can pioneer best practices in this field
- **Leveraging trustworthy AI from mature research disciplines** rather than treating AI as novel technology products, emphasising the fundamental role of researchers in building trust
- **Defining transparency frameworks** that address both how models work and the assurance processes that provide confidence in AI systems
- **Ensuring compliance with privacy legislation** whilst maintaining transparency, particularly regarding the Privacy Act 2020 requirements for handling personal information
- **Developing contextual transparency approaches** that consider model structure, deployment scenarios, data sensitivity, and impact significance

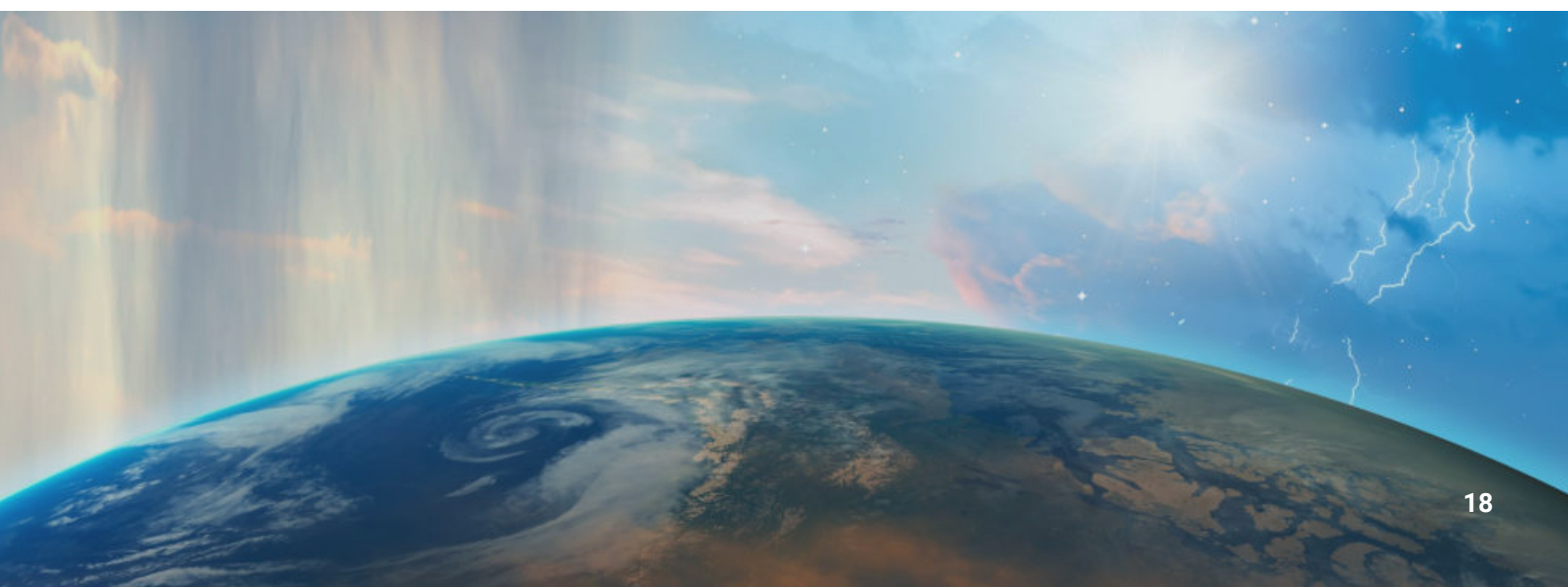
CASE STUDY: UNVEILING CLIMATE DRIVERS

Transparency and explainability are central challenges in applying machine learning to climate science, where models must not only perform well but also provide insight into the underlying physical processes driving change. Traditional global climate models (GCMs) offer interpretability through well-understood physical equations, yet they are computationally intensive and limited in scalability. In contrast, machine learning models can efficiently handle vast and complex climate datasets but often operate as “black boxes,” making it difficult to discern why a model makes particular predictions. This opacity hinders trust and limits the models’ usefulness for scientific discovery and policy decisions that demand clear causal understanding.

The DUO framework [14], developed in a collaboration between the University of Auckland and the National Institute of Water and Atmospheric Research (NIWA), directly addresses this tension by introducing mechanisms for transparent and interpretable analysis of evolving climate data. Rather than focusing solely on predictive accuracy, DUO decomposes the contributions of individual features and their interactions, revealing which variables are most influential

in shaping outcomes. Crucially, it does so in a dynamic context, identifying how the relevance of these features shifts as the data distribution changes over time. This capability is particularly important for climate studies, where drivers such as humidity, atmospheric pressure, and temperature interact in complex, evolving ways.

By quantifying and visualising the changing importance of climate drivers, DUO enhances explainability in a domain that traditionally struggles with interpretability. Its application to Aotearoa New Zealand’s station-level temperature data from 1980 to 2020 illustrates how transparent machine learning methods can yield scientifically meaningful insights — identifying specific humidity, geopotential height, and air temperature at higher pressure levels as dominant, yet temporally variable, contributors to temperature variability. Such transparency not only improves our confidence in model predictions but also provides actionable understanding for climate adaptation and mitigation. In this way, DUO exemplifies how explainable machine learning can bridge the gap between data-driven modeling and scientific interpretation.



4

ROBUSTNESS, SECURITY & SAFETY

The OECD's principle of robustness, security, and safety mandates that AI systems must operate reliably under adversarial conditions whilst mitigating risks such as data breaches, algorithmic bias, or unintended harm. This principle requires AI systems to be secure, resilient, and safe, with protections against cyberattacks, system failures, and unintended consequences. In Aotearoa New Zealand's context, this principle aligns with the nation's strategic focus on safe AI innovation through frameworks such as the Government Chief Digital Officer's Responsible AI Guidance for the Public Service [2], which prioritises security-by-design and accountability in public-sector AI projects.

Under this principle, consider AI as part of civic infrastructure, which frames expectations for reliability, safe failure, and continuous stewardship across the system lifecycle. In sensitive domains, smaller, fine-tuned, domain-specific models can narrow the attack surface, ease rigorous testing,

and temper energy use. It helps to keep in mind both technical and procedural layers: techniques for adversarial and poisoning resilience, jailbreak resistance, provenance and uncertainty signalling; alongside validated datasets, change control, audit trails, role-based access, and ongoing monitoring.

Context matters in Aotearoa. On-shore, sovereign compute can protect privacy, intellectual property, and Māori data sovereignty, and avoid opaque vendor dependencies. Co-designing safety controls with domain experts and communities, and aligning practice with tikanga and Te Tiriti, supports culturally legitimate operation. Before deployment, attention to the environment-specific stress, edge, and misuse conditions, plus readiness for incident response can clarify system limits. Making assurance artifacts visible (testing results, known limits, monitoring plans) helps operators, regulators, and the public verify responsible operation.



PATHWAY TO IMPLEMENTATION

Implementation of this principle requires embedding rigorous testing protocols, vulnerability assessments, and lifecycle risk management into AI development workflows. Three core pathways guide this implementation:

ALGORITHMIC ROBUSTNESS

Researchers must advance techniques to fortify AI models against failures, adversarial attacks, and jailbreaking risks, particularly for systems handling sensitive data such as health and forensics applications.

INFRASTRUCTURE SECURITY

Domestic high performance computing (HPC) providers such as PHF Science, NeSI, and NIWA offer platforms to host sensitive AI workloads locally, avoiding offshore vendor lock-in and ensuring compliance with Aotearoa New Zealand's Data Protection Principles.



CYBERSECURITY INNOVATION

Proactive identification of AI-specific vulnerabilities, including training data poisoning and model inversion attacks, is critical. Collaboration across the e-research system enables testing within secure, on-premises high-performance computing environments.

RESEARCH IMPLICATIONS

This principle establishes several critical research imperatives for Aotearoa New Zealand's AI community:

- **Advancing algorithmic robustness techniques** to protect AI models against adversarial attacks and system failures
- **Developing cybersecurity innovations** for AI-specific vulnerabilities and threats
- **Adopting OECD-aligned standards** to ensure interoperability whilst addressing Aotearoa New Zealand-specific needs, including protecting indigenous data sovereignty
- **Leveraging domestic HPC infrastructure opportunities** to maintain data sovereignty and compliance with national data protection principles
- **Embedding rigorous testing protocols** throughout AI development workflows
- **Conducting vulnerability assessments** as standard practice for AI system deployment
- **Implementing lifecycle risk management** to address ongoing security challenges
- **Ensuring alignment with international safety standards** whilst respecting Māori data sovereignty principles such as Te Mana Raraunga

CASE STUDY: PHF SCIENCE'S SECURE ON-PREMISES AI LAB

PHF Science's on-premises HPC infrastructure exemplifies OECD-aligned security practices through its generative AI service deployment. The organisation has implemented a multi-faceted AI platform hosted entirely on-premises and powered by open-source LLMs, consisting of both a general-purpose conversational generative AI chat system and several LLM-specialised services for data interrogation, visualisation, knowledge base management, and automated report generation [15].

This infrastructure enables PHF Science staff to automate routine tasks and enhance productivity whilst experimenting with AI capabilities with lower risk to privacy, intellectual property, or data sovereignty mandates. This approach proves crucial for organisations handling extremely sensitive data, as PHF Science deals with population health data and forensic evidence analysis.

By hosting generative AI models and bioinformatic pipelines — such as DNA

sequencing for forensic analysis — entirely within Aotearoa New Zealand jurisdiction, PHF Science ensures complete data sovereignty whilst mitigating intellectual property and privacy risks associated with offshore cloud and model providers. Sensitive health data processed through PHF Science's LLMs remains under the governance of a Public Research Organisation (PRO) with strict public-good mandates, eliminating exposure to third-party vulnerabilities whilst enabling scalable, secure AI innovation.

The platform's design demonstrates how domestic HPC ecosystems can deliver comprehensive AI capabilities whilst maintaining compliance with OECD principles. This framework positions New Zealand researchers to lead in ethical AI development whilst capitalising on local infrastructure strengths, successfully aligning global safety standards with Māori data sovereignty principles and public-sector priorities. PHF Science's model provides a blueprint for other organisations requiring secure AI deployment for sensitive applications.



5

ACCOUNTABILITY

The OECD's accountability principle mandates that AI actors should be accountable for the proper functioning of AI systems and for the respect of the first four principles, based on their roles, the context, and consistent with the state-of-the-art. This principle requires AI actors to design, install, and monitor processes that include documenting AI system decisions, conducting or allowing auditing, and providing adequate response to risks and redress mechanisms where justified. In Aotearoa New Zealand's context, accountability ensures that AI use within both public and private sectors is subject to oversight by accountable humans with appropriate authority and capability at every stage.

Accountability in this context is operational, contextual, and verifiable. Responsibilities should be delineated across the lifecycle so decisions are attributable, with records of model choices, data constraints, evaluations, and known failure modes enabling reconstruction, audit, and redress. Oversight ought to be risk-based, light where impacts are limited, and stronger where consequences are high while preserving clear human authority over

consequential uses. Because confident misuse is already occurring, accountability benefits from peer validation and fact-checking pathways that challenge outputs before they drive actions. Cultural legitimacy is integral: co-design, epistemic humility, and respect for Māori data sovereignty help prevent generic, sanitised outputs and ensure affected communities retain agency in governance and disclosure practices. Smaller, domain-specific systems, and sovereign, on-premise infrastructure can enhance traceability, privacy, and assurance, particularly for sensitive health, justice, and environmental data. Treating AI as civic infrastructure reframes accountability as continuous stewardship monitoring for drift, documenting changes, stress-testing real operating conditions, and preparing for incidents with clear triggers, roles, and remediation. Avoiding conflicts of interest by separating delivery incentives from safety assurance, and communicating capabilities, limits, and monitoring plans in audience-appropriate ways, turns accountability from principle into practice while sustaining innovation under due care.



PATHWAY TO IMPLEMENTATION

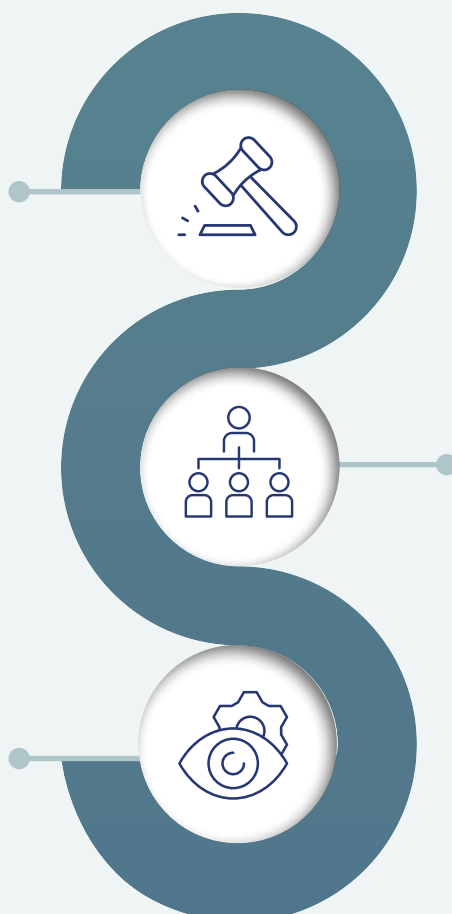
Implementation of this principle requires establishing comprehensive accountability frameworks across three core pathways:

LEGAL AND REGULATORY FRAMEWORKS

Development of laws and regulations that assign clear responsibility for AI-related decisions and impacts, ensuring compliance with Aotearoa New Zealand's existing legal framework whilst accommodating AI-specific considerations.

CONTINUOUS MONITORING AND ASSURANCE

Ensuring agency AI capabilities keep pace with technological changes to maintain strong understanding of AI systems and their limitations, whilst implementing systematic risk management approaches throughout the AI lifecycle.



ORGANISATIONAL GOVERNANCE

Encouraging organisations to adopt governance structures that oversee AI ethics and compliance, including the application of relevant regulatory frameworks, reporting mechanisms, auditing processes, and independent reviews.

RESEARCH IMPLICATIONS

This principle establishes several critical research imperatives for Aotearoa New Zealand's AI community:

- **Studying the effectiveness of accountability measures in AI governance** to identify best practices and areas for improvement
- **Proposing frameworks that delineate the responsibilities of AI developers and users** across different contexts and applications
- **Evaluating the role of audits and assessments in maintaining AI accountability** throughout the system lifecycle
- **Developing pro-innovation approaches** that ensure accountability measures do not hinder appetite for developing innovative solutions nor adopting them
- **Researching documentation and traceability mechanisms** that enable transparent AI decision-making processes
- **Investigating redress mechanisms** that provide adequate response to AI-related risks and adverse outcomes
- **Creating frameworks for human oversight** that maintain accountability whilst leveraging AI capabilities effectively
- **Developing risk management methodologies** that can be systematically applied across diverse AI applications

CASE STUDY:

LUMI™ DRUG SCAN AND PHF SCIENCE'S RESPONSIBLE AI FRAMEWORK – DEMONSTRATING AI ACCOUNTABILITY IN FORENSIC SCIENCE

Lumi™ Drug Scan represents a ground-breaking AI-based drug detection system developed through a partnership between New Zealand Police and PHF Science's forensic laboratory, combining cutting-edge technology with practical field application. Utilising machine learning models and a virtual drug reference library built from real-world forensic laboratory data, Lumi™ can scan through clear packaging to detect a broad range of illicit drug substances without requiring officers to handle potentially dangerous unknown substances. This innovative co-design approach, which involved extensive testing and validation with frontline police throughout every development stage from ideation to roll-out, ensures the system meets both scientific reliability standards and practical police usability requirements.

The development and deployment of Lumi™ Drug Scan was guided by PHF Science's comprehensive Responsible AI Framework (RAIF), which weaves together five core principles: Explainability, Fairness, Safety, Trust, and Ethics [16]. This framework consists of three key components: A Guidelines document describing purpose and process, a detailed Questionnaire defining critical considerations across project lifecycles, and a Register to document AI activities. Throughout Lumi™ Drug Scan's development, PHF Science systematically applied the RAIF Questionnaire across all five project stages – design, development, piloting, implementation, and monitoring. This structured approach ensured that key ethical considerations including data privacy, algorithmic transparency,

fairness, and human oversight were thoroughly addressed. The framework's emphasis on incorporating diverse expertise from ethics, data sovereignty, data science, accreditation, and end-user engagement helped create a comprehensive risk assessment and mitigation strategy tailored specifically for forensic science applications.

The implementation of Lumi™ Drug Scan through PHF Science's Responsible AI Framework demonstrates strong alignment with the OECD AI Principle of Accountability, which requires organisations to be responsible for the proper functioning of AI systems and respect for human rights and democratic values. The systematic completion of the RAIF Questionnaire throughout the project lifecycle created comprehensive documentation that supports transparency and accountability to multiple stakeholders, including PHF Science leadership, New Zealand Police, and the broader justice system. This documentation approach directly supports the accountability principle by providing clear evidence of responsible development practices, risk assessments, and mitigation strategies. Furthermore, the framework's emphasis on ongoing monitoring and its integration with international guidelines such as INTERPOL's Responsible AI Framework and the EU AI Act demonstrates PHF Science's commitment to maintaining accountability standards that extend beyond national boundaries. The resulting trust and confidence expressed by stakeholders illustrates how structured responsible AI practices can successfully balance innovation with accountability in high-stakes forensic applications. PHF Science's model provides a blueprint for other organisations requiring secure AI deployment for sensitive applications.

Conclusion

The Role of AI Researchers for implementing OECD AI Principles in Aotearoa New Zealand

AI research serves as a cornerstone of Aotearoa New Zealand's capacity to responsibly develop and leverage artificial intelligence technologies whilst aligning with both national values and international standards. The AI research community occupies a pivotal position in driving inclusive growth, ensuring AI technologies contribute to sustainable development and social wellbeing through evidence-based approaches that respect Te Tiriti o Waitangi and Māori data sovereignty principles.

Researchers play a fundamental role in developing AI systems that embody human-centred values and fairness, actively working to prevent algorithmic bias whilst supporting the protection of human rights. Through research into transparency and explainability, particularly in explainable AI (XAI), Aotearoa New Zealand researchers ensure AI systems remain comprehensible to the public, fostering confidence in their deployment across high-impact sectors including healthcare, governance, and finance.

The research community's contribution extends to advancing AI's robustness, security, and safety capabilities, protecting critical national infrastructure whilst safeguarding against potential adversarial threats and system failures. Equally important is researchers' role in supporting AI system accountability through collaboration with policymakers, informing the development of regulatory frameworks that clearly assign responsibilities for AI outcomes and ensuring both developers and users remain accountable for their actions.

Furthermore, researchers drive implementation of the OECD's recommendations for national policies and international cooperation, strengthening AI research and development to foster innovation whilst embedding ethical standards into technological advancement. This includes building human capacity for labour market transformation, fostering inclusive AI ecosystems, providing evidence for enabling policy environments, and facilitating international cooperation for trustworthy AI development.

AI RESEARCH NEXUS PANEL

Recommendations

The AI Research Nexus panel on 26 June 2025 (www.ainz.ai) convened leading researchers and practitioners to examine how the OECD AI Principles can be translated into concrete practice in Aotearoa New Zealand. Panelists included Gill Dobbie (University of Auckland), Markus Luczak-Roesch (Victoria University of Wellington), Mengjie Zhang (Victoria University of Wellington), and James Williams (University of Canterbury).

The discussion foregrounded engineering methods for trust, governing AI as civic infrastructure, culturally grounded approaches that respect Māori data sovereignty, and targeted investment in domain-specific, interpretable systems that deliver measurable public value. The following recommendations were discussed:



BUILD A NEW ZEALAND MODEL OF “AI ENGINEERING”

Move from abstract commitments to engineered practices: define applicability domains to clarify where models work and fail; integrate bias detection, mitigation, and adversarial-attack detection into pipelines; and require rigorous, reproducible evaluation. Share assurance artifacts (data lineage, evaluation protocols, change logs) across universities, PROs, government, and industry to avoid silos and enable audits, redress, and safe iteration. This approach draws on longstanding software engineering lessons adapted for AI, with emphasis on local context and use-bounded guarantees.



TREAT AI AS CIVIC INFRASTRUCTURE

Frame AI as part of democratic and social infrastructure rather than a purely commercial technology. Ensure public-interest research access to platforms and data, support independent scrutiny of system impacts, and enact enabling regulation so safety testing, impact assessment, and accountability are possible from the outset. This stance explicitly avoids repeating past mistakes where platform access disappeared and public researchers were locked out of studying real-world effects.



PRIORITISE SMALL, DOMAIN-SPECIFIC, INTERPRETABLE MODELS & SOVEREIGN INFRASTRUCTURE

Focus national effort where Aotearoa New Zealand has unique strengths — primary industries, climate and environmental science, health, and te reo Māori using fine-tuning and hybrid (symbolic plus statistical) methods rather than competing on frontier scale. Smaller, interpretable models can be more robust, faster to deploy, and easier to validate for local standards. For sensitive data, prefer domestic HPC or on-prem deployments to enhance privacy, traceability, and data sovereignty while avoiding offshore vendor lock-in.

4

MAKE TRANSPARENCY FIT-FOR-PURPOSE

Tailor explanations to stakeholder roles: operational diagnostics for engineers and operators, legal reasoning and evidence trails for regulators and courts, and plain-language capabilities/limits and recourse options for the public. Avoid decorative explainability that obscures detail; align disclosures with how decisions are actually made and contested. Transparency should clarify both how a model works and the assurance processes that confer confidence in its use.

5

EMBED CULTURAL LEGITIMACY & MĀORI DATA SOVEREIGNTY

Practice epistemic humility and co-design governance with affected communities, ensuring Māori data sovereignty is respected in collection, training, deployment, and disclosure. Address risks of “sanitised” generative outputs that lack nuance and authority by involving indigenous expertise, setting appropriate data controls, and designing culturally appropriate evaluation and redress pathways. Create and hold space beyond university settings (e.g., marae-based engagement) for community agency.

6

STRENGTHEN ACCOUNTABILITY WITH RISK-BASED OVERSIGHT & VALIDATION

Apply light but real, risk-based regulation; preserve clear human authority over high-stakes decisions; and establish peer-review/fact-checking pathways before model outputs drive consequential actions. Build lifecycle controls: monitor for model and data drift, manage changes with approvals and rollback plans, stress-test in real conditions, and maintain documented incident and redress procedures. Treat accountability as continuous stewardship, not one-off compliance.

7

DEMOCRATISE INNOVATION VIA OPEN SOURCE, PARTNERSHIPS & TALENT PATHWAYS

Leverage open-source tools and shared repositories, and pair them with deep domain expertise through industry-university partnerships and student placements. This enables rapid, context-specific solutions that large platforms overlook or cannot efficiently tailor. National collaboration at Aotearoa New Zealand’s scale can accelerate diffusion of methods and uplift capability across sectors.

8

TARGET “SMALL WINS” WITH CASCADING IMPACT IN HEALTH & PUBLIC SERVICES

Pursue process-level improvements where small data and “boring AI” can free workforce capacity, reduce administrative burden, and improve equity of access while also exploring designs for future health and public systems. Validate tools rigorously before deployment (e.g., automated consults), ensure transparency about data sources and accuracy, and align gains with redeployment strategies that strengthen frontline services rather than displace them.

CONFERENCE PANEL DISCUSSION:

Implementing Strategy & Principles in Aotearoa

At the AI Researchers Association Annual conference on 18 November 2025 (www.ainz.ai), at a plenary panel called “Translating OECD AI Principles into research action: challenges and opportunities for AI researchers” panelists examined the practical implications of the OECD principles and MBIE’s National AI Strategy within the constraints of Aotearoa New Zealand’s current strategic landscape. The discussion highlighted several critical areas where researchers must actively interpret and operationalise these high-level principles to fit the local context.

BALANCING ADOPTION WITH INNOVATION

The panel addressed the tension within MBIE’s National AI Strategy, which characterises Aotearoa New Zealand primarily as a “sophisticated adopter” or “fast follower.” While acknowledging the economic logic of this approach, the panelists argued that strictly adhering to an adoption-only mindset risks relegating the research community to a passive role. To truly uphold the OECD principle of **Inclusive Growth**, the ecosystem must do more than import tools; it must “self-organise” to build sovereign capability. The consensus was that researchers should not wait for top-down government directives but instead use the OECD principles as a pre-agreed ethical baseline to drive their own independent, innovation-led strategies.

CONTEXTUALISED FAIRNESS & SOVEREIGNTY

A significant portion of the discussion focused on the **Human-Centred Values and Fairness** principle. The panel identified a critical risk in relying on off-the-shelf bias detection and safety tools, which are often trained on US-centric demographics and fail to detect bias relevant to Aotearoa New Zealand. Operationalising fairness in Aotearoa New Zealand requires the development of “contextualised bias detection” frameworks that specifically account for Māori, Pasifika, and Asian communities. Furthermore, the conversation on sovereignty expanded beyond data to include “Algorithmic Sovereignty” — ensuring that the decision-making logic of AI models, not just the underlying data, aligns with local laws and Tikanga.

THE EVOLVING MANDATE OF THE RESEARCHER

Finally, the panel redefined the role of the AI researcher in this new era. With Aotearoa New Zealand identified as having high levels of “AI nervousness” despite high organisational adoption, the researcher’s mandate must expand beyond academic publication. The panel proposed that modern AI researchers must act as **Builders** (solving specific local problems), **Auditors** (ensuring long-term accountability of deployed systems), and **Translators** (demystifying AI for the public). Education was identified as the single most effective intervention researchers can offer to bridge the gap between workforce anxiety and the technical reality of an AI-enabled future.

References

1. [Approach to Work on Artificial Intelligence](#). Ministry of Business, Innovation & Employment (2024)
2. [Responsible AI Guidance for the Public Service: GenAI](#). Government Chief Digital Officer (GCDO), February 2025
3. [Aotearoa New Zealand Artificial Intelligence: A Strategic Approach](#). AI Researchers Association (2024)
4. [OECD Principles on Artificial Intelligence](#). Organisation for Economic Co-operation and Development (OECD) (2019)
5. [AI Blueprint for Aotearoa](#). AI Forum (2024)
6. [New Zealand's AI Strategy: Investing with confidence. Accelerating Private Sector AI Adoption and Innovation](#). Ministry of Business, Innovation & Employment (2025)
7. [Te Mana Raraunga](#), Te Mana Raraunga Māori Data Sovereignty Network (2025)
8. [Guidelines for the best-practice use of generative artificial intelligence in research in Aotearoa New Zealand](#) (Section 3.2.1), Royal Society Te Apārangi (2025)
9. [Sustainable Development](#). United Nations (2025)
10. [AI for social good: unlocking the opportunity for positive impact](#), Nenad Tomašev et al., Nature Communications, 11:2468 (2020)
11. [Seoul Ministerial Statement for advancing AI safety, innovation and inclusivity: AI Seoul Summit 2024](#). UK Government (2024)
12. [Keeping a close eye from space on methane emissions](#). Otago Daily Times (2024)
13. Yogarajan, V., Dobbie, G., & Keegan, T. T. (2024). [Debiasing large language models: research opportunities](#). *Journal of the Royal Society of New Zealand*, 55(2), 372–395
14. Chen, Bowen, Dobbie, Gillian, Rampal, Neelesh and Koh, Yun Sing (2024). [Unveiling Climate Drivers via Feature Importance Shift Analysis in New Zealand](#). *Proceedings of the ACM Web Conference 2024*
15. [Institute of Environmental Science and Research](#). (2025). *ESR Annual Report 2024–25*, p. 19
16. Janet Stacey, Rachel Fleming, Dion Sheppard, Jan Sheppard, Gillian Dobbie, Deepak Karunakaran (2025). [A responsible artificial intelligence framework for forensic science](#). *Forensic Science International*, Volume 375

AI

