Scoping an AI assurance regime

Andrew Jackson, former senior public servant, developed policy and approach for several national assurance systems, worked on cognitive systems in the 2000s and more recently on the use of AI in the transport sector. Policy Hub Te Herenga Waka Victoria University of Wellington

Professor Gill Dobbie, Professor of computer science focusing on machine learning and adversarial attacks, University of Auckland

Hema Sridhar, Senior Research Fellow focusing on governance, technology and Foresight, Koi Tū: The Centre for Informed Futures

Matt Farrington, Senior Lawyer, Te Herenga Waka Victoria University of Wellington

Contents

Introduction

Purpose

The purpose of this report is to provide advice to the Department of Internal Affairs (DIA) and central agencies on the scope of and approach to the development of an Artificial Intelligence Assurance Framework (AIAF) for New Zealand's public sector.

Approach

The starting point for the work was a review of a selection of international models of AI assurance together with a review of New Zealand's existing guidelines, toolkits, policy instruments and laws which are already in place to provide assurance of the use of digital technology.

This was followed by a series of conversations with academics and policy officials with expertise in technology and governance to discuss the consequent issues that would need to be addressed in the development of an effective approach to assurance.

A senior level academic oversight group reviewed and commented on this draft report.

Recommended scope of the work to develop the AIAF

The scope of the work to consider the need for and design of an AI assurance approach for New Zealand is set out below in a series of structured questions. The report which follows provides an initial assessment of the issues which needs to be considered for each of these questions based on a review of international approaches and a review of existing digital assurance mechanisms in NZ.

Recommended scope questions for the development of the assurance framework and regime

Do we need additional measures to provide assurance for the use of Artificial Intelligence in the public sector?

1. What should the goals of an AI assurance approach be?
2. Do existing mechanisms in New Zealand cover all the goals?
3. Where assurance systems are already in place to achieve those goals, does AI create technology or other specific issues which justify a separate assurance regime? If so, what are the technology or other specific issues and why do they justify an additional assurance regime?
4. Should it be a framework (co ordinating assurance activities) or a regime (setting out the principles and approach to be applied)?

If we need additional measures what should those be?

1. If an additional assurance framework or regime is needed
    a. what should its scope be?
    b. does it need to be sector specific or designed around use cases?
    c. if so, what are the categories of use case?
    d. should it be designed to assure compliance of technology or compliance of the outcomes of the use of the technology?
    e. how should compliance be assured – what are the tests?
    f. how should compliance be assured – who should apply those tests and be accountable?

What should the goals of an AI assurance approach be?

The review of existing guidelines and rules in place, indicates some of the possible goals for AI assurance and also which of these goals are already covered by existing guidelines or rules. The table below provides a high-level view of some the existing digital assurance mechanisms, the full review of the mechanisms is included in the separate written report and associated excel spreadsheet. Key findings from that work are considered under the scope questions.

One important issue not explicitly covered below is the treatment of mātauranga Māori. It is important that there is specific and separate consideration of this issue in the assurance regime as a key question under the category of "does the way the AI is being used meet social expectations". This, as with the other social tests, has to be balanced against the benefits of use such as ensuring AI is used to support learning and maintenance of te reo Māori and mātauranga Māori.

Table 1: indication of what is and what is not covered in some of the existing digital guidelines

	Are effective controls in place				Is it efficient		Does it meet laws and social expectations					Is it effective				Is it safe
	Governance	Human oversight	Capability	Risk framework	Benefits	Sustainability – energy use	Ethics	Human Rights	Privacy	Transparency	Te Tiriti	Quality Data	Bias	Accuracy	Reliability	Security
Data																
Information																
Algorithm charter																
AI toolkit																
Initial advice on Generative AI																

Key: Covered | Partial coverage | Not covered

Ethics, Te Tiriti and human rights must be carefully considered in the development of the assurance framework

Human rights

The AI assurance approach will need to meet NZ human rights law and international expectations. It is unlawful to discriminate against an individual based on a range of grounds including, ethical belief, ethnicity, age and sex.

Ethics

Ethics are what an individual considers are morally good and morally wrong and, as a consequence, provides the normative principles of behaviour. The AI assurance framework should provide assurance that the use of AI is socially, ethically acceptable. The ethics of individuals varies, develops in relation to the use of a new technology and generally evolves over time.

The process for the development of the AI assurance framework will need to include work to understand and respond to the range of ethics in relation to the use of AI. It will also need to be designed to enable the assurance framework to respond to changing social expectations.

Te Tiriti

All public sector uses of AI should respect Te Tiriti and uses of Māori data. Principles should be developed which support this. One summarised example of a list of principles is

Principle 1: Tino Rangatiratanga Māori should be engaged at an early stage to co-develop uses of AI and ensure that Māori data is stored appropriately.
Principle 2: Equity AI systems should achieve equity outcomes for Māori (individuals and collectively) across the life course and contribute to Māori development.
Principle 3: Active Protection use of Māori data in AI development requires free, prior, and informed consent (FPIC) with robust procedures in place to prevent biases or predictions that stigmatise or harm Māori
Principle 4: Mana Whakahaere Effective and appropriate stewardship or kaitiakitanga over AI systems is required.
Principle 5: Mana Motuhake Tikanga (practices) are followed throughout AI development and deployment, with Māori deciding what data and data uses are allowed
Principle 6: Tapu/Noa No AI will be culturally unsafe or break the rules of Tapu and Noa.

Source: Karaitiana Taiuru; http://www.taiuru.maori.nz/AI-Principles

Careful consideration is needed to ensure that the right balance is struck between ensuring Māori benefit from use of AI (principle 2) whilst recognising Māori data sovereignty.

Goals for an assurance regime

The first stage in the development of an assurance scheme is to agree the goals for the scheme.

Is it, for example, to achieve a good economic outcome, is it to improve the efficiency of public services or is it to achieve a good social outcome. The follow-on question is, what would the assurance regime need to test to achieve that outcome.

Conversations with officials suggest that there are three high level outcomes: to

1. Maintain public confidence and assurance that the use of AI meets public expectations;
2. Improve public sector performance, by providing confidence to the public sector to deploy AI where it can safely enhance public services; and
3. Ensure compliance with relevant legislation – such as the Privacy Act

The consequential tests for the assurance regime to achieve those outcomes might be summarized as assessing:

1. Do the benefits outweigh the costs - is there public value which supports use?
2. Will use be effective – can the AI perform the task as well as or better than humans?
3. Are we confident that use of AI in this instance will meet laws and social expectations?
4. Are we confident that use of AI in this instance will be safe and secure?
5. Have the appropriate controls been put into place to be confident that the AI use in this case will achieve what is set out in the previous 4 points?

Slide 5 shows the possible range of issues which can be considered under these 5 areas for the assurance regime to test.

The outcomes sought and the tests will also set the bar against which the success of the assurance approach can be measured.

Not covered by the assurance regime

It is important to recognize the wider issues created by the public sector use of AI.

The deployment of AI to deliver public services will allow the public service to offer new services which were not previously possible. This could have wider social implications and effects.

The additional capabilities of AI will create the opportunity to change processes and deliver efficiencies, which could have profound implications for the size, capabilities and structure of the public sector.

These are not issues which the assurance regime would address.

What we can learn from international models and experience

We looked at the approach taken in 6 other nations. Details from that work are contained in the separate more detailed written report. There is no common approach to assurance of AI use by the public sector. This demonstrates that there is no established best practice in this field and that nations are still learning what the best approach is. The differences in approach are also influenced by: different strategic national goals, with some emphasising the economic opportunities and others seeking to manage social risks; and the nature of the assurance systems already in place. It is noteworthy that over time they have developed to be more AI specific, supporting the argument that there are specific risks and skills needed to manage those risks not in place in existing frameworks. Some have developed specific approaches from existing general AI frameworks. The UK has a general framework covering AI whereas the US has developed a separate framework to cover generative AI. It is important we learn from this significant body of work that is already in place.

Commonalities

1. Whatever the regime, a growing practice is to have a senior person in each government department with responsibility for the use of AI in that department
2. They use a tick box approach to support assessment of the AI for a particular use
3. They addressed whole of life cycle for developing solutions (at each step of a project they would consider whether it met expected principles)
4. They developed standards supported by existing standards bodies
5. All had an ethics model to support assessment of whether it met social expectations

	Are effective controls in place	Do the benefits outweigh costs	How well does it work	Does it meet social expectations	Is it safe and secure
NSW - all AI					
UK					
US					
Estonia					
Finland					
Canada					

Key: Covered | Not covered

Table 2: indicating coverage of international examples of AI assurance

Do existing mechanisms in New Zealand cover all of the issues we are seeking to assure?

Reliance in the existing mechanisms is unlikely to be sufficient because:

1. There is a complex mosaic of mechanisms guidelines and regulations, without a specific AI assurance framework it is unlikely all would be considered. A single assurance framework will simplify the assurance task and increase confidence.
2. A specific regime will demonstrate that the public sector is taking seriously public concerns about the use of this technology
3. An assurance regime that is easy to use and gives confidence to departmental chief executives that the risks associated with use are being well managed is key to rapid deployment – this will particularly be the case for smaller departments with less resource
4. It is a rapidly emerging technology requiring specific expertise to understand and manage the risk and capture the opportunities

Table 3: indicating what is and what is not covered by existing digital assurance frameworks

Category of test	Examples of what it tests for	Is it addressed by existing laws and frameworks
Are effective controls in place?	Human oversight, risk frameworks	Yes, almost universally.
	Capability, governance structures	Mixed. No comprehensive governance or capability frameworks or recommendations.
Is it efficient?	Benefits, sustainability and energy use	Mixed. No comprehensive methodologies for assessing benefits or sustainability.
Does it meet societal expectations?	Ethics, human rights, data, privacy, transparency and Te Tiriti	Yes, almost universally. Though it covers these, further consideration of these issues and rationalisation will be important for the assurance framework.
Is it effective?	Bias, accuracy and reliability	Very good coverage, however not universal. Again, rationalisation may be beneficial.
Is it safe?	Security	Often not expressly addressed beyond reference to existing ISM / cloud computing .

Should it be an assurance framework or a regime?

Definitions

Assurance is defined as

> "...an objective examination of evidence for the purpose of providing an independent assessment of governance, risk management, and control processes to achieve a specified aim."

An assurance framework is a structured means of identifying and mapping the mechanisms for assurance and co-ordinating them to best effect.

An assurance regime is the approach which will be taken to deliver assurance. It is the set of implicit or explicit principles, norms, rules, and decision-making procedures to regulate the operation of an activity.

The review has found a sophisticated landscape of existing assurance approaches and associated governance systems and bodies. It is essential that the AI assurance framework considers these. However, as described in the previous slide specific guidelines and rules are needed to provide the assurance sought for artificial intelligence.

Options are a new regime that integrates with existing regimes, or consolidation of regimes into a single consistent approach.

Irrespective of formal definitions of framework and regime, care would need to be taken in the choice of the title for the approach to ensure that it receives support.

International models included both regime and framework ie rules and co-ordination– but the word "regime" was not used. Various wording is used internationally including framework, guidelines and regulations.

What should the scope of the AI assurance framework and regime be?

The scope could range from covering all forms of Artificial Intelligence through to a more focussed regime for generative AI, accompanied by a framework to co-ordinate that regime with the existing assurance processes for the use of AI more broadly.

Key issues to consider in deciding what the scope should be are:

1. The speed at which the technology is moving – so need to include a broad scope to ensure it remains relevant.

2. Ideal is to have a simple framework that is easy for the user – most of whom will not be able to distinguish the form of AI that is underpinning the technology they want to use.

3. Public expectations are that all forms should be covered.

4. There are still many additional opportunities to use predictive hence the value in having a regime that will give public confidence for use of predictive and generative AI.

5. Ensuring it matches the responsibilities and needs of the GCDO.

6. Predictive and generative AI need different forms of oversight. For instance, we know how to identify bias in predictive, but we do not know how to do that in LLMs. The algorithm charter covers predictive, could that be relied on to deal with predictive and have a separate approach for generative that is co-ordinated with that?

Definitions

Artificial Intelligence is defined as a machine's ability to perform one or more of the cognitive functions we usually associate with human minds. Cognition includes the capability to give attention to, perception, memory, learning language, problem solving, planning, reasoning, motivation and decision making.

Predictive AI analyses historic data in order to predict a future outcome based on probabilities.

Generative AI is trained on large amounts of data to build up an understanding of patterns within that data. It can then generate novel answers to questions based on its understanding of the patterns.

Should the regime be sector specific or designed around use cases?

There are certainly sector specific issues, for example the use of AI in the education sector. There is also some sector specific regulation relating to the use of artificial intelligence. For example, there is specific law relating to the use of algorithms for wine. However, the outcomes and tests to be applied to ensure those outcomes are met are sector neutral and the approach and level of assurance needed is more closely related to the way the AI will be used and the risks of that type of use.

Including sector specific assessment in an assurance regime developed and governed centrally would require deep understanding across all sector. It would seem better to develop a generic approach based on use cases and risk and for that assurance approach to include a question for those in the sector to consider whether there were any sector specific principles or regulations which need to be considered.

Very different approaches and levels of assurance will however be needed depending on the use case. For example, the assurance only needs to consider privacy legislation if the AI will have access to private data. It is noted, and the assurance approach will need to identify and classify, specific carve outs – such as the use of AI by the security services as seen in the EU AI Act.

It is noted that international models did not consider different use cases. Instead, the NSW approach for example considered risk of individual projects Further time will be needed to ensure appropriate identification and classification of use cases and associated risks.

Use cases and example

Internal administrative – support writing policy
Internal decisions - recruitment decisions
Public advisory – advice on applying for a passport
Public decision – court decisions
Service delivery – robotic surgery
Procurement of AI - AI software and things with in built AI
Regulating sectoral use of AI – deciding if ship AI systems are safe
Fraud detection - ACC claims
Security, emergency services, crisis response
Self-built AI systems – AI to summarise responses to public consultation

It is important to note that there are other ways that use cases can be categorised and different language for the categories (see the written report for another framing). If a use case approach is adopted – a key area for further investigation in the development of the assurance approach is what is the right framing. Noting it will need to be flexible as new types of use arise.

How should compliance be assured – what should the tests be?

There are two questions here – what should the tests be and who should carry out those tests. This and the next slide cover the first question and slide 15 the second question.

Two different approaches were taken to the tests in international models. Some sought assurance in the design of the technology and some in the outcome of the use of the technology. For example – do we look at the design of the AI to assess whether it will deliver unbiased results, or should we look at the outcomes of the use of the AI to see whether the outputs are unbiased. Is the right test dependent on the use and are there instances when we need to both tests?

For instance, it would not be possible to test all outcomes for procurement of AI for internal administrative support role inside a department. In this instance, the test would need to focus on whether the design of the AI met with the social expectations. While logical, the challenge would be building capability in departments and in fact within New Zealand to be able to assess the products of the providers. The EU Act deals with this by putting the responsibility to meet expectations on the seller. New Zealand does not have the scale of the EU to police compliance in this way. Procuring for use in European embassies may provide opportunity to leverage on the EU legislation? There are local suppliers who can deliver predictive and some other models which would be easier to manage for compliance.

It should be noted international assurance regimes focussed on managing risk for specific projects. If the NZ model is also considering the benefits as well as risks, the assurance approach will need to be integrated in with existing business case processes. It will need to consider the appropriate timing for the assurance test to recognise that for "2 stage" business cases it would be important to have assurance tests at both stages. Given there are existing mechanisms to assess for benefits there is also the question of whether the assurance regime can carve out that question and instead be targeted at the risks of proposals that have strong and approved business cases. Though there would be value in collecting information on benefits to demonstrate the proposed outcomes of the regime are being achieved.

The approach to assessing the risks could be designed in an efficient way by tying the approaches to level of risk as well as the use case. This would depend on the creation and adoption of a risk framework.

Low risk uses might be supported with training and guidance for the individual user.

Medium and high risk uses could require parallel human AI trials to establish capabilities and guard rails for use cases.

A valuable backstop to support the outcome of maintaining public confidence would be to establish a public register of all AI uses and performance of that AI. Also, a requirement that the public sector declare for individual services if an AI is supporting delivery of that public service and how.

How should compliance be assured – what should the tests be?

If the decision is made to test for outcomes, then the assurance scheme will need to include performance evaluation.

Some international schemes lead with the idea of performance evaluation. The UK toolkit (Section 4.1) leads with 'Gathering qualitative and quantitative data on how an AI system functions, to ensure that it performs as intended. This might include information about performance, functionality, and potential impacts in different contexts.' We need to consider whether we follow this approach.

If outcome based, the assurance scheme should include quantitative measures

- the UK framework is up-front about that. Quantitative assessment methods are mentioned both in Section 4.1 (on goals) and in Section 4.2 (on methods). People contemplating using a generative AI tool should know something specific, like 'it works on 94% of our test user queries'. These numbers are useful (a) in deciding if a new version works better, and (b) in doing detailed risk assessment work.
- The UK framework mentions 'benchmarking' AI systems. For generative AI systems that is a common way of evaluating, but it is very expensive to develop a benchmark test set; it is probably not feasible for NZ government departments, given financial pressures. Evaluating performance with human judgements might be a more practical method for us.
- Qualitative evaluations are useful too. The UK framework mentions both qualitative and quantitative methods. We might think which methods are best for which types of assessment. Quantitative is valuable for performance evaluation and red-teaming, but maybe other questions can be handled qualitatively.
- Note that for 'predictive' AI models, that learn to map A onto B from examples, quantitative evaluations are the accepted way to evaluate.

How should compliance be assured – who should carry out the tests?

What should the structure be for assurance?
The three shields model communicates the levels at which assurance can be applied:

Level 1 – guidelines and rules to be applied by the user
Level 2 – management oversight of use against the guidelines and rules
Level 3- independent assurance

And for AI case, there is also a "Level 0 "– assurance provided by the supplier if the AI is bought.

Within this framework there are also different approaches to level 2 and 3 assurance. This can either be based on a risk framework that requires assessment at level 2 or 3 of uses that are higher risk – this can be pre use, after use, or both. The second approach is to take an audit-based approach – for level 2 and 3 to assess the adequacy of the processes in place at the lower levels to ensure uses meet the tests. International models use a risk-based approach for initial use then periodic auditing.

Key considerations to consider in the design are:

1. Individuals' assessment and tolerance for risk varies
2. Need for clarity of who will be accountable for the decisions at each level and whether there should be consequences for non-compliance
3. Departments will be better placed to assess sector specific risks
4. The costs of different models (lower cost if level 2 given to existing roles, balanced against the risk of fitting in a new task with existing responsibilities and capability to deliver)
5. The need to ensure the approach is adequately resourced
6. Capability needed to ensure the assurance system works effectively (this will be a particular challenge for the smaller agencies)
7. The importance of capturing and sharing learning across the system
8. Value of a central approach to support consistency of approach and public confidence
9. A central enabling approach would allow concentration of limited resource and rapid and wide-spread adoption

Sharing good practice will support good outcomes

There are also emerging examples of good practice of the way AI is deployed to mitigate risks that are important to share and consider as general expectations in use.

For example, the ACC is using AI to assess claims. If the AI is in favour of the applicant, then the AI makes the decision, if against the applicant it is then referred to a human and the human assesses the case. This increases speed of decisions and efficiency of the process whilst maintaining public confidence.

An alternative model for other uses would be to allow the user to opt in for a lower cost AI supported process or to opt for a higher cost human process (this would of course depend on the relative costs of the two approaches).

Approach to engagement in the development of the assurance approach

Objectives of engagement

It is crucial that the assurance approach is developed and implemented in consultation with the key stakeholders. The key objectives are as follows to :

1. ensure that the public and key stakeholders have confidence and trust that there are adequate measures in place for the responsible use of AI across government.
2. gather feedback to ensure the assurance addresses the wide range of uses and scenarios and any potential gaps can be identified.
3. understand and agree the best approach to roll out of the assurance
4. identify and manage risks in the interfaces between agencies as they use AI
5. ensure there is consistency in the application of the assurance approach across various government agencies as well as to help agencies be open and transparent in their use of AI.

It will also be important to consider how to manage adversarial responses to use.

Approach to engagement

He Ara Waiora provides key principles which might underpin engagements:

•**Kotahitanga** – working in an aligned, coordinated way

•**Tikanga** – making decisions in accordance with the right values and processes, including in partnership with the Treaty partner

•**Whanaungatanga** – fostering strong relationships through kinship and/or shared experience that provide a shared sense of wellbeing

•**Manaakitanga** – enhancing the mana of others through a process of showing proper care and respect

•**Tiakitanga** – guardianship, stewardship (e.g. of the environment, particular taonga or other important processes and systems).

Key stakeholder groups

There are several different stakeholder groups that should be engaged in the development and deployment of the assurance approach:

- **Public sector**: All government agencies at all levels will be impacted by the Assurance approach.
- **Private sector**: Any businesses (ranging from multinational to SMEs, CRIs or universities) who are currently or likely to interface with the government on AI tools and applications through procurement opportunities, grants or government funding.
- **Community groups and civil society organisations**: Groups who are specifically focussed on ethics, human rights and social license aspect of AI and its use in society.
- **Māori and Pacific peoples as well as other diversity groups**: To ensure that the Assurance approach would adequately address any issues with the deployment of AI and the use of data.
- **The public** to understand the public's views on the appropriate use AI by the public sector to ensure there is social license. The assurance framework may include the expectation of public engagement before and in the approach to utliisation of AI for high-risk projects.

A more detailed analysis of the roles and responsibilities of the different stakeholder groups is included in the supporting written report.

Approach to deployment of the assurance framework

In addition to designing and winning the support of departments to the approach to assurance, consideration also needs to be given to the deployment of the regime. It needs to balance the outcome to build public confidence with the outcome to see rapid deployment to improve public services (if these outcomes are supported) in considering the approach. Rapid deployment will support the outcome of maximising public benefit through the use of AI, on the other hand careful, transparent and measured deployment will ensure public confidence is maintained in increasing use AI by the public sector.

Approaches that should be considered are:

1. Departments, identify the use cases where the risks are low – start and learn from the experience in testing for assurance and deploying the AI.

2. Take a phased approach to the application and form of the assurance framework, pick a few pilots and adapt the assurance framework in response to what we learn. If a more open approach is adopted encouraging all departments to use AI and run them through the framework it will be important to learn from and develop the assurance framework over time.

3. It should be noted that digital technology is usually tested for low medium and high-risk cases in the first instance to ensure maximum learning from the use of that technology. And in assessing the risk it will be important to consider the complexity of integration with other systems.

4. It will also be important to consider "grandparenting" arrangements. Will the assurance regime when implemented need to be applied retrospectively to existing uses of AI or just to new uses?

5. It is important that attention is also given to what future capabilities will emerge from AI to ensure that the assurance approach remains current and is not always behind the technologies which are available. One approach here would be to establish a group of experts who can advise on emerging capabilities. This would allow rapid capture of new opportunities for public benefit as well as ensuring the assurance regime is effective at managing emerging risks.

The authors acknowledge the input of key academics and policy officials in the development of this report

The authors would like to acknowledge the guidance and valuable insights from the officials who supported this work and the senior academics who looked at and commented on the draft report.

Professor Ali Knott is an expert in cognitive systems and AI and the ethics of use of AI, Te Herenga Waka Victoria University of Wellington
Professor Jonathan Boston has worked in the public and academic sectors his expertise includes public management and social policy Te Herenga Waka Victoria University of Wellington
Dr Karaitiana Taiuru is a leading authority and a highly accomplished visionary Māori technology ethicist specialising in Māori rights with AI, Māori Data Sovereignty and Governance with emerging digital technologies
Professor Karl Lofgren. Expertise includes electronic government and service delivery, Head of School of Government Te Herenga Waka Victoria University of Wellington Head of
Professor Markus Luczak-Roesch. Expertise includes the dynamics of complex systems and digital tools to augment human intelligence Chair of complexity science at Te Herenga Waka Victoria University of Wellington
Dr Simon McCallum. Expertise includes developments and capabilities of emerging AI technology. Senior lecturer computer science Te Herenga Waka Victoria University of Wellington

Adrienne Moor DIA
Kathleen Farrelly DIA
Kelly Miller DIA
Michael Daubs DIA
Miran Milosevic DIA