Te Kāwanatanga o Aotearoa
New Zealand Government

Algorithm impact assessment user guide

Algorithm Charter for Aotearoa New Zealand

December 2023

Contents

Purpose

This Algorithm Impact Assessment User Guide forms part of the Algorithm Impact Assessment (**AIA**) process and documentation prepared by Stats NZ to help government agencies meet their commitments under the Algorithm Charter for Aotearoa New Zealand (the **Charter**).

This User Guide explains the AIA process and provides guidance on how to complete the AIA Questionnaire. It is designed to support those working on algorithm projects with explanations, key considerations, case studies and links to additional materials that may be helpful. The AIA process is designed to facilitate informed decision-making about the benefits and risks of government use of algorithms.

Like the Charter, the ultimate aim of the AIA process is to support safe and value-creating innovation by agencies. Adopting a responsible approach to the development and use of algorithms and AI systems is a key contributor to innovation rather than something that stifles or blocks it. That's why the AIA process takes a risk-based approach intended to strike the right balance between ensuring agencies can use algorithms to provide better services, while still maintaining the trust and confidence of New Zealanders.

Summary of key points

Conducting an Algorithm Impact Assessment will enable agencies to identify, assess and document any potential risks and harms of algorithms so they are in a better position to address them.

Following an introduction outlining the AIA process, this AIA Guide describes a best practice approach to the issues raised in the AIA Questionnaire and helps you answer the questions in the AIA Questionnaire. That approach includes explanations, guidance, case studies, risk mitigation techniques and further reading suggestions for each of the following areas:

- AIA details
- Project information
- Potential best and worst-case scenarios
- Governance and human oversight
- Partnership with Māori
- Data
- Privacy
- Unfair outcomes
- Algorithm development, procurement and monitoring
- Safety, security and reliability
- Community engagement
- Transparency and explainability.

Acknowledgements

Stats NZ would like to thank Frith Tweedie from Simply Privacy for creating the Algorithm Impact Assessment process, designing the questionnaires, creating the report template, and compiling this user guide. Your skills and expertise have been invaluable.

We would also like to thank Katrine Evans (Government Chief Privacy Officer), Simon Anastasiadis (Principal Analyst, Stats NZ), Professor Alistair Knott (School of Engineering and Computer Science, Victoria University of Wellington), Joy Liddicoat (President of InternetNZ), the Office of the Privacy Commissioner, and Daniel Lensen (General Manager Client & Business Intelligence at Ministry of Social Development). Your insightful feedback has helped shaped these assessment tools.

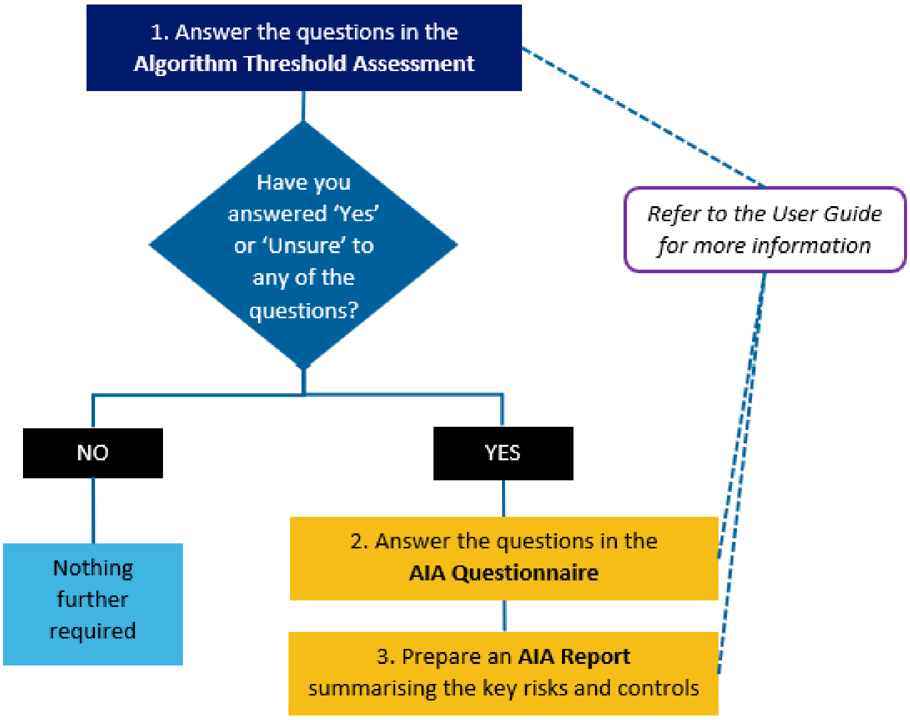
About the Algorithm Impact Assessment process

This section looks at what the AIA process entails, why it's important, when it applies, who should be involved, suggestions on how best to conduct the process and some key governing principles.

Overview of the AIA process

There are four components to the AIA process.

- **Algorithm Threshold Assessment**: an initial light-touch assessment to determine whether an algorithm presents a higher risk requiring a more in-depth assessment using the AIA Questionnaire. This assessment should be completed at the planning or design stage of a new or different algorithm.
- **AIA Questionnaire**: a series of questions about the algorithm and its possible impact, including how the algorithm works, how it will help achieve the defined objectives, what governance is in place and how it measures up against Charter commitments. This questionnaire should be completed in conjunction with this User Guide.
- **AIA User Guide** (this document): provides explanations, clarifications and case studies to support the completion of the AIA Questionnaire and help identify key risks and considerations for the AIA Report.
- **AIA Report**: articulates and summarises the key risks and controls identified in the questionnaire and serves as a record of impact assessment to support decision making.


1. Answer the questions in the **Algorithm Threshold Assessment**
Have you answered 'Yes' or 'Unsure' to any of the questions?
NO
YES
Nothing further required
2. Answer the questions in the **AIA Questionnaire**
3. Prepare an **AIA Report** summarising the key risks and controls
Refer to the User Guide for more information

The Algorithm Charter commitments

The Charter comprises six key commitments by government agencies to demonstrate their understanding that decisions made using algorithms impact people in New Zealand. Those commitments are:

- Transparency
- Partnership
- People
- Data
- Privacy, ethics and human rights
- Human oversight.

See Appendix 1 for more information.

Every agency is unique

The AIA documentation adopts a best practice approach to satisfying the Charter commitments, recognising that each agency will need to tailor the process and the ultimate risk assessments in a way that is appropriate for its own context, risk profile and role in society.

As such, each agency is free to adopt these documents - or aspects of them - as best suits their needs.

Agencies that already have assessments in place covering many of the issues raised in the AIA process may only wish to borrow certain aspects from this documentation to supplement their own processes.

Others without anything similar in place may need to adopt the full process.

A note on terminology

Please see the Glossary at the end of this User Guide for the key terms used throughout the AIA documentation – including the Algorithm Threshold Assessment - and their meaning within this context.

Although your algorithm project may involve more than one algorithm, the AIA documents refer to ‘an algorithm’ in the singular throughout for simplicity and consistency.

Similarly, while the AIA process should be conducted as early as possible, it can still be used for algorithms already in use. You will need to adjust the future tense references accordingly.

Using this guide

The User guide should be read and used by:

- designers, developers, data scientists and others working with data and on the algorithm in the context of the relevant project; and
- relevant subject-matter experts (for example, project managers; policy, privacy, ethics and legal advisers) and those in key decision-making roles.

This guidance will inevitably develop over time as technology evolves and project teams trial the process and discover its strengths and limitations. Ongoing engagement with Māori and other communities and groups may also prompt further changes. It is therefore intended that this User Guide and the related AIA materials will be reviewed and updated on a regular basis.

We recommend ensuring you refer to the most recent version of the User Guide whenever you work on an AIA Questionnaire.

Why is the AIA process necessary?

Algorithms play an essential role in the work of New Zealand's public sector by helping to streamline processes, improve efficiency and productivity, enable the faster delivery of more effective services and support innovation. Algorithms can also help to deliver new, innovative, and well-targeted policies to achieve government aims.

However, it's well established that the opportunities afforded by new and evolving technologies can also introduce potential risk and harm. That includes challenges associated with accuracy, bias and a lack of transparency, explainability, reliability and accountability. At a societal level, poorly governed algorithmic and AI systems can amplify inequality, undermine democracy and threaten both privacy and security.

As AI tools incorporating algorithms become increasingly sophisticated and commonplace – including the explosion of Generative AI models such as ChatGPT – it is more important than ever that government agencies approach these technologies with due care and diligence.

A responsible approach to the development and use of data, algorithms, and AI, with clear accountability and risk management standards across the algorithm lifecycle will not only help produce better quality algorithms, it will also lead to higher levels of trust and help maintain the social licence of the agency using them.

This perspective is inherent in the Algorithm Charter, a set of commitments made by government agency signatories to carefully manage how algorithms are used. The Charter was released in July 2020 to increase public confidence and visibility around the use of algorithms within the public sector.

What is an algorithm?

The Charter does not provide a formal definition of ‘algorithm’ because the risks and benefits associated with algorithms tend to be contextual and are largely unrelated to the type of algorithm being used. Very simple algorithms can result in just as much benefit or harm as more complex ones depending on the context, focus and intended recipients of the outputs.

A broad definition of ‘algorithm’ is included in the Glossary as a high-level guide only, along with examples of the definitions adopted by some agencies.

The Algorithm Threshold Assessment will also help you to determine if the Charter applies and a full AIA is required. It contains a short list of questions designed to ‘weed out’ those algorithms that are unlikely to have a material impact on people and a corresponding risk of harm.

When does the AIA process apply?

Algorithm Threshold Assessments should be completed at the planning stage of a new or different algorithm. Where required, the AIA Questionnaire should also be initiated at an early stage, noting it may need to be worked on and updated throughout the development process.

Using the Algorithm Threshold Assessment

The full AIA process – that is, using the AIA Questionnaire to produce an AIA Report - is designed to apply to higher risk algorithms only. It is not intended that every business rule or process used by agencies will be captured by this process.

The Algorithm Threshold Assessment replaces the assessment using the Risk Matrix set out in the Algorithm Charter. You should use the screening questions in the Algorithm Threshold Assessment to identify if the algorithm in question is likely to automate, aid, replace or inform any operational or policy decisions that are likely to have a ‘material impact’ on individuals, communities, or other groups.

The Algorithm Threshold Assessment also asks whether the algorithm uses any sensitive personal information or any form of artificial intelligence, including Generative AI, machine learning, facial recognition or other biometrics likely to have a ‘material impact’ on any individuals, communities or other groups.

What is a ‘material impact’?

A ‘**material impact**’ means a decision that could reasonably be expected to affect the rights, opportunities or access to critical resources or services of individuals, communities or other groups in a real and potentially negative or harmful way, or that could similarly influence a decision-making process with public effect.

Determining whether a material impact is likely to occur requires consideration of various factors, including:

- the nature of the potential impact
- who might be affected
- how long that impact might last
- how significant or severe the impact or consequences are likely to be
- the likelihood of the risk occurring.

For example, decisions affecting people's legal, economic, procedural or substantive rights (including those relating to the administration of justice or democratic processes, human rights, and privacy rights) or those that could impact eligibility for, or access to services (including education, social welfare, health, housing, ACC or immigration services) are likely to have a material impact, while decisions likely to result in only minor or highly unlikely impacts will not be considered material.

Examples of algorithms likely to have a material impact and therefore provoke a "Yes" answer to the screening questions (triggering the need for completion of the AIA Questionnaire) include:

- A machine learning algorithm that generates a score for citizens that is used to help a government department determine their eligibility for benefits
- The use of facial recognition technology to compare and identify citizens for security purposes.

Examples of algorithms that are not likely to have a material impact include:

- An algorithm being used by an agency to transform image to text (for example, to digitise handwritten documents) as part of an archiving process
- An automated scheduling tool which sends out internal diary invites from a mailbox.

These examples demonstrate that the purpose for which an algorithm will be used and the wider context are key to determining the likelihood of some form of material impact. For example, the outcome is likely to be different for the image to text algorithm above if it was being used to digitise paper application forms for a government service. In that context, poor performance of the algorithm on some handwriting styles could influence success rates for individual applicants.

Accordingly, when completing the Algorithm Threshold Assessment, you should discuss whether a material impact is likely with a multi-disciplinary group, including those with privacy, legal and risk management expertise.

If you tick "Yes" or "Unsure" to one or more of the questions in the Algorithm Threshold Assessment, you should complete the AIA Questionnaire and produce an AIA Report. The general rule of thumb is, if in doubt, please complete the full process.

Other risk management frameworks still apply

Please note that the AIA documentation is not a complete list of all requirements for algorithm projects. Project teams should always ensure they comply with their agency-specific legal obligations, processes, risk management frameworks, policy requirements, and governance mechanisms.

Note that simply completing the AIA process does not mean that an algorithm has an acceptable risk profile and is fine to proceed. The final output of this process – the Algorithm Report – should identify and clearly articulate the relevant harms, risks and mitigants so appropriate decision makers in your agency can take accountability for the algorithm and whether it is acceptable or not to use it as intended.

Privacy considerations

Questions relating to the collection, use and sharing of data are relevant to both privacy law and ethical considerations associated with algorithms.

For example, data provenance and usage, accuracy, transparency, reliability, security and accountability are all key privacy considerations where personal information is involved – and you will find those core privacy concepts are embedded across all sections of the AIA Questionnaire and this User Guide.

Failure to appropriately identify and manage privacy risks can result in harm to individuals and creates legal and reputational risk for your agency. Accordingly, where the Project or the algorithm involves personal information, you must engage with your Privacy or Legal team to understand whether a Privacy Impact Assessment (PIA) is required either before or in parallel with the AIA process.

While the AIA addresses some algorithm-specific privacy considerations in the "Privacy" section towards the end of the document, it does not replace the need for a PIA.

Generative AI

Generative AI uses prompts or questions to generate text or images that closely resemble human-created content. Generative AI works by matching user prompts to patterns in training data and probabilistically "filling in the blank." ChatGPT is the most well-known, free, example of Generative AI.

By enabling people to quickly and easily create new content, Generative AI offers many public service benefits, including greater productivity and faster, more efficient innovation.

However, it may also present a range of risks, including in relation to privacy, accuracy, security, Māori Data Governance, procurement and intellectual property rights. Generative AI algorithms also enable and scale the rapid creation of harmful content such as misinformation, revenge porn and media content intended to sow social discord. Moreover, because AI algorithms reflect the values and assumptions of their creators, they can perpetuate conscious and unconscious biases that lead to exclusion or discrimination.

As Generative AI is being integrated into many commonly used public sector tools, it is something that public servants will need to take pro-active steps to manage. Any public sector use should align with the 'Initial advice on Generative Artificial Intelligence in the public service' produced by data, digital, procurement, privacy and cyber security system leaders across the New Zealand public service (Joint System Leads tactical guidance on Generative AI). That includes developing an appropriate Generative AI policy for your agency and fully assessing and actively managing risks, including through the use of privacy impact and security risk assessments.

An AIA will be appropriate where the Generative AI is likely to have a material impact on any individuals, communities or other groups. This means that, for example, using a Generative

AI tool like ChatGPT - or versions embedded in office software products - for relatively low-risk tasks such as helping to write an email will not require completion of the full AIA process.

However, where the use of Generative AI tools could reasonably be expected to significantly affect individuals, communities or other groups, particularly, where they are being used in circumstances where inaccuracy, bias, mis/disinformation present real risks, an AIA Questionnaire and AIA Report should be completed.

Please also see the Privacy Commissioner's expectations around Generative AI and how to manage the potential privacy risks associated with using such tools.

Who should be involved in the AIA process?

A multi-disciplinary team

The AIA process is best completed by involving a diverse and multi-disciplinary range of inputs from people with a wide range of knowledge, skills and experience. While the precise combination will depend on the nature of the algorithm, the agency and the Purpose, that could include the following roles.

Business owner(s)	Project Manager	Business Analyst(s)
Data scientists and engineers	Designers and developers of algorithms and AI systems	Systems architect
Privacy & ethics advisers	Legal advisers	Security advisers
Māori data sovereignty experts	Procurement specialists (where applicable)	Staff who will use the algorithm
Community advocates		

Where a co-design approach is being used, you should also include community advocates.

Support from external advisers may also be appropriate, particularly for higher risk algorithms or where internal capability is not available.

Input from key internal advisers is critical

Subject to the nature of the data and algorithm to be used, it's likely you will need to engage with your agency's internal privacy, security, and legal advisers. External advice may be needed for more high-risk projects or where internal capability is unavailable.

Please ensure you engage with these teams as early as possible to establish when and how they will need to be involved.

Privacy team

Your privacy team should be consulted to ensure the privacy impacts of algorithms using or processing personal information - or that otherwise impact individuals' privacy rights - are identified, assessed, and mitigated.

A Privacy Impact Assessment (PIA) may be required, which will help identify actions and approvals required under privacy legislation and policy, including the Privacy Act, the Data Protection and Use Policy and other internal data-related policies applicable within your agency.

Legal team

When completing an AIA, you should consult your legal team to identify and address any legal risks arising from the development, procurement or use of the proposed algorithm and wider system(s). Consultations should begin at the concept stage of a Project, prior to development or procurement.

The nature of the legal risks will depend on the design of the system (for example, the training data or model used), the context of the proposed usage and the nature of the outputs.

Security team

Your security team can advise on how best to ensure the algorithm and related data are kept secure and free from external or internal misuse or attack. They can perform a Security Impact Assessment, which is assumed to be completed in association with the questions in section *10 Safety, security and reliability* of the AIA Questionnaire.

Broader engagement and collaboration

Engagement and collaboration with Māori and other communities impacted by the algorithm is key. See the further discussion in the sections *Partnership with Māori* and *Community engagement* relating to the Charter's 'Partnership' and 'People' commitments.

The 'People' Charter commitment requires active engagement with people, communities and groups with an interest in algorithms and consultation with those likely to be impacted by their use.

The principle of *Mahitahitanga* in the Data Protection and Use Policy (DPUP) is also relevant here. See Appendix 2 for more information. This means working together to create and share valuable knowledge and involves including a wide range of people in projects or activities that collect or use people's information. The principle also advocates for working with iwi and other Māori groups as Te Tiriti o Waitangi partners, ensuring they are involved in decisions about data and information issues that affect them.

Who should complete the documentation?

Every agency and every AIA is different so there are no hard and fast rules as to who is best placed to be the author of the AIA documentation. However, the following roles would usually be involved in completing the documentation.

Algorithm Threshold Assessment and AIA Questionnaire: Typically completed by someone in the Business Owner's team, who will take responsibility for co-ordinating the multi-disciplinary information gathering and collation of results. Please refer to this User Guide for support when completing those documents.

AIA Report: Someone with an understanding of risk and an ability to clearly identify and articulate the relevant risks, harms and mitigants should prepare the AIA Report, which is a critical tool for empowering informed decision making. This might be someone in a Risk, Privacy or Legal role. This person should conduct the final review of the AIA Questionnaire. External support may be appropriate for particularly high-risk projects.

If there are any questions in the AIA Questionnaire that you're unable to answer, please note this down in your answers, including why (for example, because relevant information is unavailable or the question can only be answered after certain things have occurred). Unanswered questions provide an answer in themselves and play a role in the risk profile of the algorithm in question. This should be reflected in the AIA Report.

Plain language please

Please ensure you use plain, clear and simple language when you populate the AIA documentation, avoiding jargon and technical terms where possible.

If the jargon or technical terms must be used, please clearly explain their meaning so a non-technical audience are able to understand the concepts.

How should we conduct the AIA process?

Each agency will have its own way of doing things and its own governance and risk frameworks that will continue to apply. This User Guide does not aim to provide a prescriptive process, but rather to include some ideas on how you might want to approach the AIA process to get the best results.

Workshops and group activities

Workshops can be an effective and efficient way for multi-disciplinary groups to brainstorm issues and surface a range of different perspectives, particularly where you need to:

- define the problem, algorithm use case, purpose and overall benefits
- identify who will be impacted by the algorithm (the **Impacted People**) and the potential benefits and harms of using the algorithm in relation to each type of Stakeholder
- identify potential mitigants to the risks and harms identified for each type of Stakeholder.

There are various techniques that can be effective for brainstorming and information gathering in workshops. You many find a **Lean Canvas** approach helpful to start your design journey. This is a one-page model to guide teams through a series of defined steps, helping you to focus on the key aspects of your problem and how you might solve it in a responsible way.

While a Lean Canvas is typically used by entrepreneurs to brainstorm business models, the overall concept can be useful in the AIA context as well.

A good example of a machine learning canvas can be found in this article, which is also referenced in the Ministry of Social Development's *Data Science Guide for Operations* part of its Model Development Lifecycle at page 45.

The Open Ethics Canvas and the Ethics Canvas are other tools that may be useful.

What should we cover in workshops?

We recommend holding at least one initial workshop as early as practical to gather and discuss the following information.

Project goals and purpose (as required for section 2 of the AIA Questionnaire). That includes defining the issue you're trying to solve or goal you're trying to achieve; the nature of the proposed algorithm, what it does, how it will be used and how it will help (particularly as compared to the status quo); the key benefits, Impacted People and users; technology considerations and key dates. See the related guidance in the following section.

Identification of Impacted People and key impacts, including the potential benefits and harms or negative impacts of using the algorithm in relation to each group of Impacted People.

Identification of potential mitigants to the risks and harms identified for each group of Impacted People, as covered in section 8 of the AIA Questionnaire.

Consider best and worst-case scenarios - see Question 3 in the AIA Questionnaire. Think about what could go wrong and how you might explain the rationale behind deploying and using the algorithm to the media or a judge. Who is likely to be held accountable and will they be able to justify why the algorithm was used and how it was developed and deployed? How will they demonstrate that sufficient consideration was given to potential unfair outcomes, risks and unintended consequences? Applying this thinking at an early stage can really help crystallise your thinking around why and how you use an algorithm.

Consider holding subsequent workshops to check on progress and, as a group, identify and brainstorm emerging risks and mitigants over the course of the Project. A workshop mid-way through the Project provides an opportunity for re-assessment and re-direction if needed.

A final workshop towards the end of a Project can help to confirm the algorithm appropriately satisfies the Purpose, is adequately resourced, delivers the anticipated benefits and does not cause material harm to individuals, communities or other groups.

You may find it helpful to include a good facilitator who understands the issues can keep participants focused and ensure the necessary discussion points are covered and key information is surfaced.

Trials and Proofs of Concept

Trials, pilots and proofs of concept (POCs) can play an important role in developing and successfully deploying and using an algorithm. By adopting a smaller and narrower focus in the first instance, you can test and learn before moving to full-scale deployment.

This can help build confidence in the data and the algorithm before they "go live" and are used in ways that impact people. You might like to conduct research into the use of similar technology overseas or in other contexts, as well as running research on how the algorithm performs on a sample data set. This will also provide behavioural insights into how the algorithm may be used, which can enable trouble shooting and help you anticipate unintended consequences.

Note however that pilot projects are likely to be quite different from the development of algorithms at scale, since larger-scale projects may require time to make sure the right data is gathered, the appropriate use case is chosen and costly mistakes are not made while developing technological architecture.

It's unlikely you will need to complete the AIA process for most POCs, though that will be context dependent (it may be prudent to complete the process for particularly high profile/high risk POCs). However, you may still find the questions in the AIA materials helpful when designing, conducting and reviewing your POC. Consideration of the relevant issues is also likely to assist in getting approval from decision makers to proceed with a more formal and wide-ranging project/initiative, as well as ensuring you will already have a solid base for completing the AIA process where required.

Where can we find support?

Interim Centre for Data Ethics and Innovation

The Interim Centre for Data Ethics and Innovation (ICDEI) supports government agencies to maximise the opportunities and benefits from new and emerging uses of data, while responsibly managing potential risk and harms.

The ICDEI's role is to raise awareness and help shape a common understanding of data ethics in Aotearoa New Zealand, while building a case for a wider mandate and a scaled-up work programme over time. It will work across a wide network of people and ideas, drawing on the knowledge and expertise within that network, including the Data Ethics Advisory Group.

Where other parts of the network are already undertaking work (like this Algorithm Impact Assessment), ICDEI's role is to support, accelerate, and use the network to promote and disseminate the work.

Data Ethics Advisory Group

The Government Chief Data Steward (GCDS) has convened a Data Ethics Advisory Group (DEAG) to help maximise the opportunities and benefits from new and emerging uses of data, while responsibly managing potential risk and harms. This group will enable government agencies to test ideas, policy and proposals related to new and emerging uses of data. It will also provide advice on trends, issues, areas of concern, and areas for innovation.

The DEAG consists of members with expertise across privacy and human rights law, ethics, innovative data use and data analytics, Te Ao Māori, technology, public policy, government interests in the use of data (social, economic, and environmental), Pasifika and community representation.

More information on the group is available here.

Government Chief Privacy Officer

The Government Chief Privacy Officer (GCPO) leads an all-of-government approach to privacy to raise public sector privacy maturity and capability. The GCPO supports government agencies to meet their privacy responsibilities and improve their privacy practices.

The GCPO is responsible for providing leadership by setting the vision for privacy across government, building capability by supporting agencies to lift their capability to meet their privacy responsibilities, providing assurance on public sector privacy performance and engaging with the Office of the Privacy Commissioner and New Zealanders about privacy.

See here for more information.

Government Chief Information Security Officer

The Government Chief Information Security Officer (GCISO) is the government system lead for information security.

The role strengthens government decision-making around information security and supports a system-wide uplift in security practice.

The GCISO’s work includes coordinating the Government's approach to information security, identifying systemic risks and vulnerabilities, improving coordination between ICT operations and security roles, particularly around the digital government agenda, establishing minimum information security standards and expectations and improving support to agencies managing complex information security challenges.

See here for more information.

Office of the Privacy Commissioner

The Office of the Privacy Commissioner (OPC) works to develop and promote a culture in which personal information is protected and respected in New Zealand. It is an Independent Crown Entity that is funded by the state but which is independent of Government or Ministerial control.

OPC has a wide range of functions, including investigating complaints about breaches of privacy, building and promoting an understanding of the privacy principles, monitoring and examining the impact that technology has on privacy, developing codes of practice for specific industries or sectors, monitoring data matching programmes between government departments, inquiring into any matter where it appears that individual privacy may be affected, monitoring and enforcing compliance with the Privacy Act and reporting to government on matters affecting privacy, both domestic and international.

See here for more information.

AI Forum New Zealand

The Artificial Intelligence Forum of New Zealand (AI Forum) is a purpose-driven, not-for-profit, non-governmental organisation that is funded by members. It was founded in 2017 to bring together New Zealand's community of AI technology innovators, end users, investors, regulators, researchers, educators, entrepreneurs and the interested public to work together to find ways to use AI to help enable a prosperous, inclusive and thriving future for our nation.

The **AI Governance Working Group** was established to provide thought leadership on the responsible governance of AI in Aotearoa and develop a curated set of frameworks, tools and approaches that meet the needs of New Zealand organisations. It is producing a toolkit on AI governance that includes tools, approaches and principles to help organisations operationalise responsible AI governance. It is also developing a list of AI Governance groups and experts in New Zealand, with details coming soon.

See here for more detail.

Completing the AIA Questionnaire

1. AIA details

Why this is important

The algorithm is likely to be part of a wider piece of work, so this section captures the overall name of that wider Project, which is then used throughout the rest of the AIA documentation, where relevant.

General guidance

Enter the name, role and contact details for the key personnel involved in the Project, including the accountable Executive Sponsor (such as a Deputy Chief Executive or similar).

2. Project information

Why this is important

This section of the AIA Questionnaire captures key details about the Project to help contributors, reviewers and decision makers understand why the Project is being undertaken, what it involves and the expected impacts. This information is critical to informing subsequent questions.

General guidance

Be sure to attach or link to any relevant documents or supplementary information about the algorithm and the wider Project, such as the business case, specifications, general project documentation, system architecture diagrams, data flow maps, user interface designs, user instructions and manuals, legal advice, privacy advice, Privacy Impact Assessments, security advice.

Key considerations and risks

Problem and purpose (Questions 2.1, 2.2)

Clear identification and articulation of the issue or problem you are trying to solve is critical to ensuring you select the right algorithm or other approach to solving that issue.

When describing the problem or issue you are trying to solve, consider the following alongside the 'status quo comparison' relating to Question 2.3 discussed below under the heading *Identifying other Project information.*

- The scale of the issue

- What inequities exist
- How many people are impacted
- Specific impacts for Māori
- Current solutions and management

'**Purpose**' is a key definition used throughout the AIA process.

It refers to how and why the algorithm helps achieve the objectives of the Project in the relevant business context. This is particularly important when considering accuracy, potential biases and other unfair outcomes, which tend to be highly contextual. Throughout the AIA process, you and your Project contributors should remain focused on **why** the algorithm is being used, what you are trying to achieve and how it will help you address the issue or problem articulated above.

Identifying impacted people (Questions 2.5, 2.6)

A workshop with multi-disciplinary participants can be a good way to brainstorm and identify the full range of individuals, groups and communities who are likely to be impacted by the algorithm's use (the Impacted People). The following discussion prompts can be used to guide your discussion.

Who are the Impacted People?		
Who	Example	Discussion prompts
People who will be **directly** impacted by the algorithm	• People whose data is being processed by the algorithm • People who are the subject of a decision made by the algorithm	○ Whose data will be processed by the algorithm? ○ Who will be evaluated or monitored by the algorithm (whether by choice or otherwise)? ○ Who will the algorithm make predictions or recommendations about?
People who will be **indirectly** impacted by the algorithm (those who are affected by the algorithm in a less obvious or immediate way)	• People responsible for those who are the subject of an algorithmic decision or whose data is being used (such as parents, guardians, whānau, hapū) • Children and whānau of those who are the subject of an algorithmic decision or whose data is being used • Communities who may be affected by the algorithm but who don't use it • Society at large.	○ Who is in the vicinity of the algorithm that may be impacted by its use? ○ Who may have a significant interest based on their relationship to other Impacted People? ○ Which communities may be affected? ○ Who may be concerned about compliance or the ethical implications of the algorithm?

Who are the Impacted People?		
Who	Example	Discussion prompts
People who will **use** or **access** the algorithm.	• Staff • Members of the public where an algorithm is made directly available.	○ Who will interact with or use the algorithm? ○ How and when will they use it? ○ Who will interpret the outputs? ○ Who will manage, operate, oversee or control the algorithm? ○ Who will decide whether to use the algorithm for a particular task?

In particular, consider the potential impacts on the following groups.

- Māori
- Pasifika
- Tāngata whaikaha - people with disabilities
- Ethnic communities
- Refugees and migrants
- Whanau/families
- Tamariki/children and students
- People with mental health conditions and addictions
- Older people
- People in rural and remote areas
- People experiencing homelessness, poverty, violence and so on.
- Takatāpui - rainbow communities
- Prison populations.

Once you have identified the Impacted People, you can then identify the potential benefits and impacts of using the algorithm for each type. This information can initially be recorded in a very simple table (like the one below) before being refined for inclusion in response to questions 2.5 and 2.6 in the AIA Questionnaire.

WORKSHOP: Identifying Impacted People and potential impacts		
Impacted People	Potential benefits	Potential harms
1.		
2. [and so on…]		*(Together, the* ***Impacted People****)*

Please note: clearly identifying the Impacted People and associated benefits and harms is an essential element of the AIA process - it is critical to understanding the potential impact of the algorithm on people and not just your agency.

Identifying other project information

Please succinctly provide sufficient detail to enable someone with no familiarity with the algorithm to understand what the algorithm does, where it comes from, and its role in the wider Project.

The following prompts are designed to help you provide an appropriate level of clarity. Question numbers from the AIA Questionnaire are also listed.

- **Source** (Question 2.3): Where does the algorithm come from? Is it being developed internally or procured from a third party? Why was that particular algorithm or wider system selected? For algorithms obtained from an external supplier, please ensure comprehensive answers are provided to question 9.2 of the AIA Questionnaire.
- **Outputs** (Question 2.3): Clarify the nature of the algorithm's outputs – for example, does it produce decisions? Predictions? The creation of new content?
- **Status quo comparison** (Question 2.3): How are tasks completed currently or decisions made without the algorithm? How will the algorithm change things?
    - Explain how the benefits and risks of current practices compare to the benefits and risks of using the algorithm. What will be different, including both improvements and downsides.
    - Is the algorithm selected the best tool for the job and is it a proportionate response to the problem? What would be lost if the algorithm were not used and could any alternative approaches achieve similar results with a different/lower risk profile? Describe the potential alternative approaches and why they have been discounted.
- **Benefits** (Question 2.4): What are the potential benefits of using the algorithm as planned and the likelihood of such benefits being realised in practice? For example:

    - Delivering a better existing service or outcome
    - Delivering a new service or outcome
    - Reducing processing or delivery times
    - Generating financial efficiencies or savings
    - Enabling future innovations to existing services, or new services or outcomes.

- **Value proposition** (Question 2.4): Describe what value the algorithm is expected to deliver and to whom, both on its own and as part of the wider Project? Can a dollar figure be placed on this value or is it more intangible? Consider how success is being defined in the Project, how it will be measured and how the algorithm will contribute towards that success.
- **Values and expectations** (Question 2.4): Is use of the algorithm consistent with your agency's core values and purpose? Is it consistent with New Zealanders' expectations of your agency?
- **Artificial Intelligence** (Question 2.7): Where a form of AI is involved (noting the broad definition used in the Glossary), please describe the form of AI that is proposed. For example, does it involve machine learning, natural language processing, computer vision (for example facial recognition or other biometric technologies) or something else? It may be a combination – for example, Large Language Models typically involves both machine learning and natural language processing.

CASE STUDY: Automatic tax refunds

To address the time-intensive process of finalising a customer's tax each year, Inland Revenue implemented a new system to calculate an individual's tax position where they are reasonably confident of that person's income.

It uses an algorithm to complete a calculation on the customer's behalf and issue an immediate refund or notice of outstanding tax.

This has made the tax return process much less onerous for many people, with a high proportion of taxpayers now having to do little or nothing when their tax return is due.

Source: Algorithm-Assessment-Report-Oct-2018.pdf (data.govt.nz), page 16

- **Generative AI** (Question 2.7): If a form of generative AI such as ChatGPT is proposed, describe if it will be used to create any new data, text or content and how such material will be used and by whom.
- **Surrounding technology** (Question 2.8): Describe the extent to which the algorithm will interact with any other hardware or software systems currently in use and the nature of those systems. Consider who develops or deploys those systems and whether the way they interact with the algorithm needs to be factored into the overall AIA considerations.
- **Lifecycle dates** (Question 2.10): Please set out each of the anticipated dates for *each stage* of the algorithm's lifecycle, using estimates where firm dates are not yet known. Your responses here will help reviewers and decision makers understand the overall release timeline.
- **Team and diversity** (Question 2.11): Do the people handling the data and developing the algorithm have the necessary qualifications and expertise? What is their relationship to your agency – are they employees, contractors, consultants and so on.?

  Diverse teams are better equipped to bring a range of perspectives that can help minimise potential bias and other unfair outcomes. Ideally, project teams will consist of people with a diverse range of skills, experiences, genders, ethnicities, ages, abilities and backgrounds, particularly for the team(s) preparing the relevant data and developing the algorithm.

You should ensure there are Māori involved in the Project with a view to ensuring Māori perspectives can be embedded in the design, development, testing and implementation of the algorithm (noting Māori also have a diversity of views and no one person, whānau, hapū, iwi or other Māori organisation or community can speak for all Māori).

If there are gaps in your diversity, it's important to think about whether any particular perspectives are missing. If so, you may need to take additional steps at the community engagement stage to ensure those perspectives are incorporated into the Project.

3. Overall risk profile (potential best and worst-case scenarios)

Why this is important

This section asks you to describe the best and worst-case scenarios that could arise from using the algorithm. These questions aim to start laying the groundwork for articulation of the key risks in the Algorithm Report.

General guidance and key risks (Questions 3.1 to 3.3)

Best-case scenario

Describe the best-case scenario that could arise from use of the algorithm, including a description of:

- the key beneficiaries and how and why they will benefit. Who might miss out or be disadvantaged in this scenario?
- the nature of the public benefit and how this will be recognised.
- the likely challenges or hurdles to achieving the best-case scenario.
- the likelihood of this eventuating.

Worst-case scenarios and potential harms and risks

Please describe the worst-case scenario(s) that could arise from use of the algorithm, including a description of what this might look like both when the system works as designed or intended **and** when the system fails or doesn't work as designed or intended in some way.

CASE STUDY: Young people not in employment, education or training (NEET)

Work and Income's Youth Service (NEET), uses an algorithm to help identify school leavers at greater risk of long-term unemployment to proactively offer them qualification and training support. The algorithm considers factors shown to affect whether a young person may need support and produces risk indicator ratings to indicate the level of support that might be required. It also refers school leavers to NEET providers for assistance and determines funding.

Since 2012, a third of the more than 60,000 young people that have accepted assistance were offered the service through the algorithm. NEET has proved to be most effective for those with a high-risk rating, resulting in improved education achievements and less time on a benefit, compared with those who did not use the service.

Source: Algorithm-Assessment-Report-Oct-2018.pdf (data.govt.nz), page 14

To help you think about how these scenarios might play out in real life, for each scenario please describe the following:

- The nature of the potential harm or impact and who is most likely to be impacted, how, and why.
- How and why those harms could arise.
- For each harm, describe how *likely* it is to occur and how *severe* the impact is likely to be on the Impacted People.
- Who, if anyone, might benefit in each scenario.
- Whether and how those harms could be avoided.
- If the harms cannot be avoided, how they will be addressed.
- How those impacts will impact the Purpose.
- The media headlines that could appear and who in your organisation or within government or elsewhere is likely to be held accountable.
- Who would take responsibility for fixing the identified failures, errors or unfair outcomes.

What risks are presented to the agency as a result of these harms (for example, compliance breaches, reputational damage for the agency and/or Government as a whole, the possibility of an independent investigation being commissioned, financial implications)

Risk mitigation options

- What controls or mitigants are **already in place** to address the identified harms and risks?
- What **further controls or mitigants** may be necessary to address those harms and risks?
    - What changes are needed to actively address the potential harms, including across the algorithm lifecycle and more broadly (for example like governance, transparency, engagement).
- If those controls are implemented, how will that change the extent of risk or harm identified above?

4. Governance and human oversight

Why this is important

Governance refers to the relationships, systems and processes within and by which authority is exercised and controlled.

It is ultimately about accountability and, in the context of algorithms and particularly AI systems, involves ensuring an appropriate framework of policies, practices and processes is in place to manage and oversee the use of algorithms and associated risks to ensure the use of these tools aligns with the organisation's objectives, is developed and used responsibly and complies with applicable legal requirements. Governance of algorithms should be part of an agency's overall governance framework.

General guidance

Key components of good governance (Questions 4.1 to 4.3)

- A clear **strategy** and purpose for using algorithms and AI
- Clear **accountability** and clearly defined and documented **roles and responsibilities**:
  - for design, delivery, monitoring and risk management across the algorithm lifecycle
  - for making challenging ethical decisions. Recourse to a multi-disciplinary ethics review board or similar mechanism (whether internal or independent) can help assess ethical challenges, including how to balance possible trade-offs and navigate unclear grey areas.
- **Policies, processes and procedures** setting out the appropriate development and use of algorithms and data to drive a systematic and consistent approach (including policies addressing data governance, privacy, data retention, risk management, ongoing monitoring and audit). The Joint Systems Leads tactical guidance on Generative AI also recommends developing a policy and standards for trialling and using Generative AI.
- A **risk management framework** including approval, monitoring and audit processes.
- Appropriate **capability**, including suitably qualified and experienced people to review the AIA Questionnaire and clearly identify and articulate the key risks, potential harms and mitigants in the AIA Report. Who carries out this step should be decided by the agency.
- **Training** on appropriate use of data and algorithms.

Key considerations, risks and mitigation options

Review and audit (Question 4.4)

Record keeping facilitates reviews and the identification and rectification of issues. Higher-risk algorithms with the potential to have a significant adverse impact on people may need to be independently audited.

Forms of Generative AI such as OpenAI's ChatGPT or Google's Bard are known to confidently create and present inaccurate content, often referred to as their tendency to 'hallucinate' or make prediction errors. You will need to implement robust review and fact checking processes to ensure any inaccurate content is detected where such tools are used.

To facilitate auditing, ensure the development process and training data sources are well documented, log the algorithm's processes and maintain an appropriate audit trail for any predictions or decisions made by the algorithm.

Human review (Question 4.5)

Human review and oversight can help ensure an algorithm is performing as expected and mitigate the risks of unfair outcomes.

Your answers to Question 4.5 of the AIA Questionnaire should detail the nature of the relevant human review and oversight, including whether the algorithm will:

- replace a decision that would otherwise be made by a human
- replace any human decisions requiring judgement or discretion
- be used to assist a human decision maker make a decision.

Generally speaking, the less oversight a human can exercise over an algorithm or AI system, the greater the need for more extensive testing and stricter governance.

CASE STUDY: Visa triage

Immigration New Zealand developed a triage system that includes software to assign risk ratings to visa applications to guide the level of verification required on each application.

The algorithm does not determine whether an application is approved or declined and an Immigration Officer still assesses and determines each application.

Use of the algorithm has increased consistency across visa processing offices, improved processing times, and allowed attention to be focused on higher-risk applications. This allows staff to identify new and emerging risks and see where risks are no longer present.

Source: Algorithm Assessment Report (Internal Affairs, Stats NZ), page 17

Automation bias (Question 4.6)

While human review can help ensure an algorithm is performing as expected, this can be undermined where 'automation bias' occurs. 'Automation bias' refers to "the tendency to over-rely on automated outputs and discount other correct and relevant information" (Source: https://assets.publishing.service.gov.uk/media/5d7f6b2540f0b61ccdfa4b80/RUSI_Report_-_Algorithms_and_Bias_in_Policing.pdf, at p. 15).

The unquestioning acceptance of automated decisions and recommendations can lead to system errors being overlooked, potentially leading to harm.

You should aim to ensure an appropriate balance of machine and human decision-making and implement suitable safeguards to minimise the risks of automation bias, including:

- paying particular attention to the design of the user experience and user interface, including adding reminders in online systems for users to exercise their own judgement
- enabling human users to choose to ignore algorithmic recommendations and exercise their own reasonable discretion where appropriate
- providing staff training and developing guidelines and other awareness-raising materials about automation bias, including showing human users how the system can get things wrong and emphasising the responsibility of human users (studies such as this one show a greater sense of individual human responsibility results in greater critical engagement and more careful information processing).

Legal considerations (Question 4.7)

A range of legal obligations may apply to use of the algorithm, including under the Privacy Act 2020, the Human Rights Act 1993, the New Zealand Bill of Rights Act 1990, the Copyright Act 1994, and agency and sector specific legislation, as well as administrative and public law principles and legislation.

Please talk to your legal team to identify and address any specific legal or regulatory concerns. That is particularly important for automated decision-making algorithms, where there may be statutory powers that cannot be delegated or fettered without parliamentary approval.

Please make sure any legal advice is attached to the completed AIA Questionnaire.

Appeals and recourse (Question 4.8)

Citizens are entitled to challenge decisions made about them by government agencies. This can also help to surface inaccuracies and unfair outcomes.

The ability to challenge an algorithm-based decision is inherently connected with transparency and explainability – people need meaningful explanations about how decisions affecting them have been made.

It may help to establish a process or 'feedback loop' for Impacted People to report potential vulnerabilities, risks or unfair outcomes. You should also have a manual or similar alternative process available in case the algorithm is not performing adequately.

Guidance and training (Question 4.9)

Appropriate guidance and training should be provided to all staff who interact in a material way with the algorithm. That may include training on:

- how to understand the algorithm's outputs and decisions and how to identify and address potential errors, failures, and unfair outcomes
- automation bias risks (discussed above)

- the need to check the accuracy and reliability of outputs, particularly for Generative AI tools.

ADDITIONAL GUIDANCE

- The US National Institute of Standards and Technology (NIST) provides guidance on AI governance in its AI Risk Management Framework. The 'Govern' function sets out the requirements for governance of an AI programme, including policies and procedures, accountability structures, diversity, culture and risk management.
- The Ministry of Social Development's Model Development Lifecycle provides a governance framework, including the need for clearly defined roles and responsibilities, approval gates, the recording of risks and the ownership of both controls and risks.
- The *Policy Risk Assessment* questionnaire developed by the New Zealand Police includes a section on 'Oversight and accountability' that sets out requirements for individuals to certify that the technology can do what is claimed, the proposal has been assessed from privacy, security, Te Ao Māori and ethical perspectives, and that the technology is explainable.

5. Partnership with Māori

Why this is important

In Aotearoa New Zealand, the governance relationship between the Crown (government) and Māori is shaped by Te Tiriti o Waitangi. The Charter reflects the Crown's commitment to honour Te Tiriti o Waitangi and ensure the use of algorithms is consistent with the articles and provisions in Te Tiriti.

The articles and provisions of Te Tiriti can be summarised as (but are not limited to):

- Article 1 - good governance
- Article 2 - self-determination and active protection
- Article 3 – equal and equitable participation
- Article 4 – protection of Māori customs and beliefs

The courts and Waitangi Tribunal have described Te Tiriti generally as an exchange of solemn promises about the ongoing relationships between the Crown and Māori, including the promise to protect Māori interests and allow for Māori retention of decision-making in relation to them.

To ensure algorithm use delivers clear benefits to iwi and Māori, government agencies need to build trust and form enduring relationships with iwi and Māori to:

- understand the interests and role iwi and Māori may have in algorithm development and use
- appropriately mitigate risks and potential negative consequences
- ensure ethical use of data provided by iwi and Māori

- minimise the risk of algorithm development and use not being consistent with Te Tiriti
- meet the expectations introduced in the Public Service Act 2020, which reflect the contemporary needs and opportunities in the Māori Crown relationship.

New Zealand is also a signatory to the United Nations Declaration on the Rights of Indigenous Peoples (UNDRIP). UNDRIP is a comprehensive international human rights document on the rights of indigenous peoples. It covers a broad range of rights and freedoms, including the right to self-determination, culture and identity, and rights to education, economic development, religious customs, health and language.

General guidance

To meet the Partnership commitment in the Charter you should:

- incorporate te ao Māori perspectives into the design and use of algorithms
- ensure algorithm development and use is consistent with Te Tiriti o Waitangi
- consider how Māori data sovereignty will be maintained
- assess how algorithm use will impact iwi and Māori.

Te ao Māori acknowledges the interconnectedness and interrelationship of all living and non-living things via spiritual, cognitive, and physical lenses. This holistic approach seeks to understand the whole environment, not just parts of it. (This definition comes from Treaty of Waitangi/Te Tiriti and Māori Ethics Guidelines for: AI, Algorithms, Data and IOT.)

Māori are diverse in terms of interests, aspirations and needs. There is no one Māori world view but multiple te ao Māori perspectives.

Key considerations

Find out what your organisation already knows (Question 5.1)

Use what your organisation already has and knows. It's important to coordinate engagement activities and not duplicate what others have already done to minimise the burden on iwi and Māori.

- Find out who in your organisation has relevant relationships with iwi and Māori and seek their advice.
- Find out what relevant engagement activities have already occurred and what was learnt through those engagements.
- Seek guidance from te ao Māori, tikanga, and mātauranga Māori experts in your organisation.
- Discover what data is held by your organisation as well as previous projects and research that align with the intent of the algorithm. Employ lessons already learned.
- Find out what relationships have been established with other agencies to see what engagements and insights they may have.

Conduct a Te Tiriti analysis (Question 5.1)

A Te Tiriti analysis will help you identify how the articles and principles apply to the development and use of the algorithm in your Project. The Policy Project Toolbox collates guidance on conducting this analysis – a key resource is the Cabinet Office Circular on Te Tiriti o Waitangi Guidance.

Engage with iwi and Māori early (Question 5.1)

When engagement with iwi and Māori is needed, ensure you engage early in your project – building trust and relationships needs time and space. Be conscious of timeframes – relationships should not exist solely for the duration of a project. Decide how you will sustain these relationships to move them from extractive and transactional to enduring and reciprocal. Consider whether decision-making will be shared and, if so, how. Share what actions have resulted from the input and contribution of Te Tiriti partners.

Follow Te Arawhiti's engagement framework and guidelines to ensure you have appropriately identified Māori interests. Te Arawhiti's resources include:

- a Crown engagement with Māori framework
- Guidelines for engagement with Māori, using the engagement framework
- an Engagement Strategy Template
- principles for building closer partnerships with Māori.

The Ngā Tikanga Paihere framework may also be helpful in encouraging you to be mindful of those potentially impacted by your algorithm and aid in developing a more holistic view that includes te ao Māori perspectives.

You may also want to check the Settlement Portal – Te Haeata, an online record of Treaty settlement commitments. This portal helps agencies and settled groups search for and manage settlement commitments.

Be aware of the overwhelming demand from government on iwi and Māori to engage and consult on issues of concern. This often happens without creating the conditions for engagement and consultation to take place in a way that works for these communities. Do not contribute to the overwhelming of iwi and hapū leaders and Māori experts.

If iwi and Māori say no, it doesn't necessarily mean they're not interested. It's likely that they're already participating in many other government kaupapa, or this isn't a priority for them right now.

Co-design approach (Question 5.2)

Te Kāhui Raraunga has published a Māori-Crown Co-design Continuum which identifies three main types of Māori-Crown co-design and two other design approaches where Māori and Crown design independently:

- Mana Māori co-design
- Ōritenga co-design
- Participatory co-design
- Māori Motuhake design

- Crown Exclusive design.

The Continuum can be used as a planning tool for Māori co-design initiatives, whether they are initiated by Māori, or Māori are invited by Crown agencies to co-design.

Māori data (Question 5.3)

Māori data is defined as data that is about, from or by Māori, and any data that is connected to Māori. This includes data about population, place, culture, environment and their respective knowledge systems. (This definition comes from https://www.kahuiraraunga.io/iwidataneeds)

Māori data is not owned by any one individual, but is owned collectively by one or more whānau, hapū or iwi. Individuals' rights (including privacy rights), risks and benefits in relation to data need to be balanced with those of the groups of which they are a part. (This definition comes from https://www.temanararaunga.maori.nz/)

Māori data sovereignty recognises that Māori data should be subject to Māori governance – the right of Māori to own, control, access and possess Māori data. Māori data sovereignty supports tribal sovereignty and the realisation of Māori and iwi aspirations. (This definition comes from https://www.temanararaunga.maori.nz/)

For Māori, data is a taonga. Māori are and have been data designers, collectors and disseminators for generations. The form Māori have collected it in differs from the modern understanding of data, and the forms that data is collected and transmitted are closely interconnected with Māori mātauranga and ways of being. Data was and continues to be how Māori have continued their consciousness as Māori across time and distance. (This definition comes from https://www.kahuiraraunga.io/iwidataneeds)

Māori recognise that Māori data contains a part of the people or natural environment that the data is about, no matter how anonymised the data is.

Improved outcomes for Māori (Question 5.4)

Algorithm use should contribute to improved outcomes for Māori.

Impacts on Māori cannot be fully identified from a non-Māori perspective, so consultation with Māori is essential. (Source: Cabinet Office Circular CO(19)5)

- What is the impact on Māori? Is this different to the impact on all New Zealanders? If yes, how and why?
- Will the use of the algorithm enhance Māori wellbeing?
- Will the algorithm affect different Māori groups differently?
- What could the unintended impacts on Māori be and how do you propose to mitigate these?

Additional Guidance

Links to other helpful guidance:

- Cabinet Office Circular CO (19) 5: Te Tiriti o Waitangi / Treaty of Waitangi Guidance
- GIDA CARE Principles

- Māori Data Sovereignty
- Māori Data Governance Model
- Iwi Data Needs
- UNDRIP
- Te Haeata - the settlement portal

6. Data

Why this is important

The Charter's *Data* commitment requires signatories to make sure data is fit for purpose by understanding its limitations and identifying and managing bias.

Data is the life blood of algorithms, so it's critical to understand what training and production data will be used, how accurate and reliable it is and whether it is suitable to use in the Project.

Remember that data is a taonga (treasure) and of high value so it must be treated with respect, particularly where it relates to people and Māori in particular.

Using the right data in the right context can substantially improve decision making but the opposite is also true. A long-standing rule of data science is GIGO – Garbage in, Garbage out. Poor quality data can lead to inaccurate, unreliable and even harmful results, particularly where historical biases in data are not considered and addressed.

General guidance

To help decision makers identify and understand any potential adverse impacts before relying on algorithmic insights or decisions, please clearly describe in your AIA Questionnaire answers both the data that will be used to train the algorithm (**training data**) as well as the data that will be fed into the algorithm once it has been deployed (**production data**). Please make sure you provide answers in relation to **both types of data**.

Key considerations and risks

Data sources (Questions 6.1, 6.2 and 6.4)

Please explain the sources of your data, including who collected the information, from whom, when, where, how and for what purposes. If this information is unknown or unavailable, please be sure to state this and the reasons why it is unclear.

CASE STUDY: Skin cancer detection

Algorithms used to detect skin cancer can operate more accurately than dermatologists.

However, research has shown that algorithms trained on images taken from people with light skin tones only might not be as accurate for people with darker skins and vice versa. This can arise due to underrepresentation in training datasets as well as inconsistencies in the devices used for image acquisition and selection.

Source: Characteristics of publicly available skin cancer image datasets: a systematic review - The Lancet Digital Health

For personal information, consider whether the information was collected from individuals with their knowledge and consent. Even where consent is not be a legal requirement, you are more likely to have social licence to use that information if the relevant people have agreed to it being collected and used in the first place.

See the discussion on consent in A Path to Social Licence – Guidelines for Trusted Data Use published by the Data Futures Partnership.

For example, personal information collected many years ago for a different purpose may not be legally available for you to use depending on your Project's context and proposed use case. Please consult with your Privacy team in such situations.

When it comes to using data obtained from external parties, you should ensure you understand the source and method of data collection, particularly for training data. Sensitive personal information sourced from overseas may have consent requirements associated with collection (for example, information sourced from Australia, the UK and Europe) and you will need to have visibility of whether it was collected in accordance with the law. Again, please consult with your Privacy and Legal teams.

Data ownership (Question 6.2)

Consider who owns your training and your production data and whether you have the appropriate rights to use that data in the Project. Note that, strictly speaking, no one "owns" personal information; rather you are a steward or custodian of that data and the Privacy Act 2020 governs how it is to be handled and protected.

You should pay particular attention to the ownership of data generated using Generative AI tools, which may not be clear cut. Please consult with your Legal team around any copyright or other intellectual property questions.

Storage (Question 6.3)

Will the data be stored in New Zealand or offshore (including in overseas data centres)? If the data includes personal information, offshore storage may require compliance with IPP 12 of the Privacy Act 2020 – discuss with your Privacy team.

Relevance, sufficiency and representation (Question 6.5)

For an algorithm to be effective, its training data must be representative of the communities to which it will be applied. This is particularly important for algorithms developed externally (for example, by a supplier) or trained on overseas data.

- Consider how will you ensure the relevant datasets accurately represent the populations to which the algorithm will be applied.
- Do you need to engage on this topic with Māori, affected communities, subject matter experts and civil society groups?

Make sure you have sufficient data to achieve the Purpose. Is any additional data required to improve accuracy and reduce the risk of unfair results?

If so, where will you get it from and how will it be obtained?

Accuracy (Question 6.6)

Statistical accuracy refers to the ability to produce a correct or true value relative to a defined parameter. In the context of an algorithm, that's likely to reflect how closely an algorithm's outputs match the correct labels in test data.

As discussed in the Ministry of Social Development's (MSD) *Data Science Guide for Operations* (Data Science Guide) that forms part of its Model Development Lifecycle (MDL), the accuracy of an analytical model usually depends more on data preparation methods than on the model type or the model tuning procedures.

The Data Science Guide notes "it is inevitable that data will contain some errors", including historical, systematic and random errors. It suggests that for each data variable used, the different sources of error should be considered, mitigated and documented.

It goes on to set out some of the key decisions that must be made in preparing the data are as follows (noting MSD's comments that this isn't an exhaustive prescription of how to prepare data).

- Is the historical data relevant to the business context or are there marked differences in the business process that need to be accounted for?
- Are there inherent biases in the data, and how will these affect the model outcomes?
- What variables should be used as inputs to the model and what is the target variable?
- Are there useful proxy variables that can be constructed to represent unobservable variables? (see the discussion on proxy variables in the Unfair Outcomes section)

- Are there any errors or missing values that should be accounted for?
- What transformations need to be applied to the data?
- Are there any influential outliers?
- Is there a need to account for rare occurrences or imbalances in the data?

The *Data Science Guide for Operations* notes that, as with most of the decisions made in building an algorithm, the answers to these questions rely on understanding the business objective. Accuracy needs to be considered alongside other criteria, including transparency, ease of implementation and the potential for bias – and trade-offs between the relevant criteria may be necessary.

Where this is the case, you should ensure potential trade-offs are considered by a multi-disciplinary team (including business owners and privacy and legal representatives) in the context of the use case, the algorithm's contribution to the Purpose and the business context and any potential risks that may result from any such trade-offs.

Generative AI

Where you are proposing to use Generative AI, you should also consider and document how you will address the specific risks of so-called 'hallucinations'. That is, the tendency of Generative AI to 'make up' information or return out-of-date, biased or misleading results.

As noted in the Joint System Leads tactical guidance on Generative AI, ensuring the accuracy of AI outputs is critical. It is essential that data used to train AI tools is of high-quality, for quality outputs. Cleansing, validating and quality-assuring data can help to ensure accuracy and reliability of outputs.

See also the discussion on Performance and testing in section 9 of this User Guide, which includes examples of some appropriate metrics for measuring accuracy and other performance issues.

Risk mitigation options

Various techniques are available for documenting data used in algorithms and AI tools, such as:

- **FactSheets**: documents detailing the purpose, performance, accuracy, safety, security and provenance of data in AI systems to engender consumer trust in an AI service.
- **Datasheets for Datasets**: documenting key aspects of a dataset, including its composition, collection process and recommended uses, helping to facilitate better communication and encouraging developers to prioritise transparency and accountability.
- **Model Cards for Model Reporting**: short documents accompanying machine learning models that explain the context in which models are intended to be used, details of performance evaluation procedures and other relevant information. This helps to support transparent model reporting. See here for a user-friendly explanation.
- **Dataset Nutrition Labels**: A diagnostic framework that provides a distilled yet comprehensive overview of dataset 'ingredients' before AI model development. This can

help drive more robust data analysis practices, provide an efficient way to select the best dataset for a project's purposes and increase the overall quality of models.

- **Reward Reports**: improves the ability to analyse and monitor AI-based systems over time.

ADDITIONAL GUIDANCE

For more detail on data preparation for use in algorithms and some of the key decisions to make where preparing data, see the Ministry of Social Development's *Data Science Guide for Operations* that forms part of its Model Development Lifecycle (MDL).

The MDL is an open-source set of documents intended to be used by other agencies. It consists of a *User Guide*, a *Governance Guide* and a *Data Science Guide for Operations*, all of which aim to provide decision makers with assurance that technical, legal, ethical and Te Ao Māori opportunities and risks are managed throughout an algorithm's lifecycle.

7. Privacy

Remember to attach or link to a copy of any Privacy Impact Assessment already conducted, as well as answering the questions in the AIA.

Why this is important

As noted in the introductory section About the Algorithm Impact Assessment process, privacy considerations are embedded throughout the AIA process, including in relation to questions of data collection, quality, security, accuracy, transparency and access.

Algorithm-specific privacy issues are raised in the *Privacy* section of the AIA Questionnaire, including some that might not automatically be considered in a standard Privacy Impact Assessment (PIA).

Key considerations, risks and mitigation options (Questions 7.1 to 7.4)

Algorithms that use large volumes of personal information – or particularly sensitive personal information – may require a separate PIA.

As there may be some cross-over between the work done in each of the AIA and PIA processes, you should speak to your Privacy team at the earliest opportunity. They can help you to determine the best risk assessment processes for the Project and which information is relevant in what context to avoid duplication and ensure consistency. This is particularly important in relation to risk descriptions and controls so as to avoid confusion.

Your Privacy team can also help you to understand the key privacy risks and mitigants, ensure a Privacy by Design approach is embedded across the Project and articulate privacy considerations in workshops within the AIA process. Early-stage workshops as discussed in the Project information section can also be a helpful starting point for privacy teams to gather project information to inform a PIA.

Generative AI

Where Generative AI tools are proposed, please refer to the following New Zealand public sector guidance, which will be updated as the technology evolves.

- Guidance from the Office of the Privacy Commissioner on Generative Artificial Intelligence
- Joint System Leads tactical guidance on Generative AI

Biometric technologies

Biometric technologies enable the automatic recognition of people based on their biological or behavioural features, including their faces, eyes (iris or retina), fingerprints, voices, signatures, keystroke patterns, odours or gait.

The use of biometric technologies can present significant risks, including in relation to surveillance and profiling, bias and discrimination, and a lack of transparency and accuracy.

The Office of the Privacy Commissioner (OPC) has clearly stated that biometric information is personal information that is regulated by the Privacy Act 2020. It is currently exploring whether to establish a new biometrics Code of Practice pursuant to the Privacy Act and consultation is underway.

Project teams looking to use biometric technologies should ensure they are compliant if the Code of Practice comes into force. As always, please engage with your Privacy team at the earliest opportunity where biometric technologies are proposed as part of a Project.

ADDITIONAL GUIDANCE

- The *Kaitiakitanga* principle in the Data Protection and Use Policy includes keeping data safe and secure and respecting its value, as well as acting quickly and openly if a privacy breach occurs. See the webpage for the Kaitiakitanga principle for more details.

- The Ministry of Social Development's Privacy, Human Rights and Ethics Framework (PHRaE) is a helpful tool for identifying and addressing privacy, human rights and ethical risks. It includes guidance documents relating to personal information, data security, transparency, bias, operational analytics, partnership and automated decision-making.

8. Bias and other unfair outcomes

Why this is important

'Unfair outcomes' is the term used in the AIA Questionnaire to refer to bias, discrimination and other unfair, unintended or unexpected outcomes. These unfair outcomes may occur for a variety of reasons, including as a result of historical data that reflects cultural biases or as a consequence of how the algorithm itself is developed or used. They can often have significant impacts as discussed in this article.

Anyone subject to algorithmic decision-making by government agencies or courts in New Zealand has legal protection from discrimination under the New Zealand Bill of Rights Act 1990 and the Human Rights Act 1993.

CASE STUDY - Criminal justice

Algorithms used by criminal justice systems across the United States to predict recidivism were found to be biased against Black people. Black defendants were more likely than white ones to be incorrectly judged as having a higher risk of re-offending.

Source: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Discrimination is unfair treatment based on the protected characteristics in the Human Rights Act 1993, which include sex, race, ethnic or national origin (including nationality or citizenship), disability, age and employment status.

It's important to note that bias leading to unfair outcomes can still occur even it does not meet the requirements for discrimination under the Human Rights Act.

General guidance

Bias can be defined as a systematic difference in the treatment of certain objects, people or groups in comparison to others. This can occur due to a range of factors, including:

- pre-existing cultural, social, or institutional perspectives
- the nature of the training data (which may reflect those pre-existing perspectives and societal biases)
- engineering decisions
- the use of algorithms in unexpected contexts
- decisions and processes across an algorithm's lifecycle.

CASE STUDY - Recruitment

Amazon dropped an AI-powered recruitment tool to review CVs after it was found to penalise female job candidates.

The algorithm was trained to identify patterns in successful job applications at Amazon over the previous 10 years – the majority of whom were male. So even though the algorithm was not explicitly trained to look at gender, it taught itself to prioritise male candidates because of historical hiring biases.

Source: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

Fairness: The related but distinct concept of 'fairness' deals with more than just the absence of bias. Fairness considerations typically relate to the overall outcomes of an algorithm, including the need for fair decisions to be reasonable, to consider equality implications, to respect personal agency and to not be arbitrary.

There are multiple perceptions of fairness and it is a highly conceptual and often ambiguous concept. This can be particularly challenging for algorithms, as fairness is not a notion with absolute and binary measurement.

It is therefore critical that human developers and users of algorithms decide on an appropriate definition of fairness for each Project and the specific context. The target outcomes and their trade-offs must be specified with respect for the relevant context.

The concept of equity is also important, particularly in a health context.

What might unfair outcomes look like?

- Common categories of negative or unfair impacts or outcomes include the following types (SA TR ISO/IEC 24027:2022 Information Technology – Artificial Intelligence – Bias in AI Systems and AI aided decision making).
- Inequitable allocation of opportunities, resources, or information
- Failure to provide the same service or quality of service to all
- Reinforcement of existing societal stereotypes
- Denigrating people by being actively derogatory or offensive.
- Over or under-representation of certain groups, or even failure to represent them.

These types of harm are not mutually exclusive and an algorithm or wider Project can involve more than one type.

CASE STUDY - Benefit fraud

A Dutch court ordered the Dutch tax authority to stop using an algorithm to predict childcare benefit fraud due to breaches of human rights and privacy laws. After more than 20,000 families were wrongly accused of benefit fraud, the tax authority admitted that at least 11,000 people were singled out for special scrutiny because of their ethnic origin or dual nationality, fuelling longstanding allegations of systemic racism in the Netherlands. It was also criticised for a lack of appropriate checks and balances surrounding use of the algorithm.

- Several victims committed suicide and over a thousand children were taken into foster care.
- The Dutch privacy regulator issued €6.45m in fines, including for the "unlawful, discriminatory and therefore improper manner" in which the tax authority processed data on the dual nationality of childcare benefit applicants.
- More than €500m was set aside in compensation and the Dutch government resigned as a result of the scandal.

Source: https://techcrunch.com/2020/02/06/blackbox-welfare-fraud-detection-system-breaches-human-rights-dutch-court-rules/

Key considerations and risks

Bias is often unintentional

There are many different types of bias and they can be conscious or unconscious (that is, outside a person's conscious awareness). See the definition of 'Bias' in the Glossary for an indication of some of the different types of bias, noting this is not an exhaustive list.

Context is key (Question 8.1)

Given the many complex sources of bias and other unfair outcomes, it is not possible to completely remove bias or guarantee fairness. In fact, bias can have positive, neutral and negative effects and even may be desirable in some cases (for example, to correct for historical under-representation – see the surgery waitlist equity adjustor case study below).

CASE STUDY – Surgery waitlist equity adjustor

Ethnicity is one of five criteria used in a Te Whatu Ora algorithm to prioritise waitlisted patients for surgery. The equity adjuster waitlist tool aims to reduce inequity in the health system by considering clinical priority, time spent on the waitlist, geographic location, ethnicity (specifically Māori or Pacific people) and deprivation level. Clinical need takes precedence.

Source: https://www.hinz.org.nz/news/642771/Auckland-algorithm-improves-equity-of-waitlists.htm

What constitutes a negative effect will therefore depend on the context, business goals and the overall Purpose for using the algorithm and will need to be carefully considered by a multi-disciplinary team.

Those collaborating on an algorithmic project should clearly define and document what fairness means in that particular project, ensuring a diverse range of perspectives contribute to that definition.

Consider and document any trade-offs (Question 8.1)

Avoiding bias and achieving fairness can involve having to consider trade-offs relating to the defined Purpose and competing priorities.

It is always important to balance performance metrics against the risk of unfair outcomes - you may want to consider monitoring a set of metrics (see further discussion in the Algorithm performance, testing and monitoring section) that balances performance across several dimensions. Transparency and documentation of priorities and underlying assumptions are essential.

Bias can occur throughout the algorithm lifecycle (Question 8.2)

- **In the data:** If input datasets used to train an algorithm are not sufficiently comprehensive, up-to-date or representative of the populations to which the algorithm will be applied, this can have disproportionate impacts on any groups not properly represented in datasets.
- **In the algorithm:** Biased logic, flawed assumptions, inappropriate modelling techniques, coding errors and pattern misidentification in the design process can lead to biased outputs.
- **In usage:** Incorrect interpretation of algorithm outputs, inappropriate use of those outputs (for example, automation bias) and disregarding underlying assumptions can create bias.

Those designing and using algorithms need to be aware of the various ways in which unwanted bias can be introduced and then design, test and validate their systems to correct for potential unfair outcomes.

Proxy variable risks (Question 8.3)

Even when information that may cause discrimination is not present in a dataset, it is still possible to discriminate by using ‘proxy variables’. These are often used where relevant data is not readily available or is difficult to measure.

Difficulties can arise where the data used could represent or correlate with sensitive attributes or prohibited grounds of discrimination. The exclusion of protected characteristics from training or input data does not guarantee that outcomes will not be unfair, since other variables could serve as close proxies for those characteristics. For example, a postcode could operate as a proxy for race.

A number of otherwise benign features can combine to become a proxy from which sensitive information may be inferred. AI systems are particularly efficient at identifying underlying patterns which may correlate to a protected attribute like ethnicity, and as a result could make predictions or decisions which create the risk of bias or discrimination.

Bias and discrimination risks with biometric technologies

Biometric technologies (biometrics) have the potential to produce biased or discriminatory outputs, particularly where training data is significantly different from production data. This can also be caused by the fact that systems designed to detect physical characteristics (for example, faces, fingerprints) may need to manage a wide range of variables than if detecting a uniform object like a swipe card.

If a biometric system detects the characteristics of a certain group less accurately than others, it's likely to produce biased outcomes. This could result in discrimination against a particular group.

Risk mitigation options

You may be able to *identify* unwanted bias at various points during the development and deployment of algorithms, including when:

- internal requirements (such as objectives, strategies and purpose) and external requirements (for example, legal and regulatory requirements) are defined
- Impacted People are identified
- data sources are selected and documented.

You may be able to *address* unwanted bias when:

- decisions are made as to how best to represent the training data in features interpretable by the algorithm (that is, 'feature engineering')
- data is labelled, trained and tuned
- the algorithm is verified and validated
- the algorithm is monitored and validated after deployment.

See also the section on Algorithm performance, testing and monitoring in this User Guide for examples of algorithm performance metrics.

Various tools are available to enable transparent reporting of algorithm provenance, usage and fairness-informed evaluation – see 'Risk mitigation options' in the Data section of this User Guide.

ADDITIONAL GUIDANCE

- The Ministry of Social Development's *Data Science Guide for Operations* that forms part of its Model Development Lifecycle provides helpful practical guidance on steps to assess fairness and bias.
- The New Zealand Police's *Guidelines for Algorithm Development and Life-cycle Management* include a set of questions and metrics to assist with measuring the fairness of an algorithm.
- ISO Standard SA TR ISO/IEC 24027:2022 on *Bias in AI systems and AI aided decision making* also provides helpful practical support.

9. Algorithm development, procurement and monitoring

In-house algorithm development

Why this is important

The benefits of agencies developing algorithms in-house include greater visibility and control of training data and the algorithm itself, as well as easier oversight and clear lines of accountability. Agencies are able to develop an algorithm in line with their defined Purpose and are well positioned to clearly define and document the algorithm's technical features and performance metrics.

General guidance

Good practice includes recording the use of algorithms and AI systems in an internal inventory with accompanying information relating to its source, usage and basic technical details.

See also the discussion on open-source solutions below, which may also be relevant in projects involving the internal development of algorithms.

ADDITIONAL GUIDANCE

For more detail on algorithm selection and optimisation, see the Ministry of Social Development's *Data Science Guide for Operations* that forms part of its Model Development Lifecycle.

External procurement

Why this important

As algorithms and AI tools become increasingly sophisticated, agencies are likely to seek to procure such tools from specialist external suppliers rather than developing them in-house. Third-party AI tools, including open-source models, supplier platforms and commercial

APIs, are now commonplace and AI-as-a-Service is a growth trend involving the use of AI tools built by others in the cloud.

While there may be valuable cost savings and quality considerations that support a procurement approach to algorithms, the specific risk profile of externally sourced algorithms necessitates appropriate due diligence.

Agencies remain responsible for any harms or risks caused by third-party algorithms or AI tools.

General guidance

The questions in the AIA Questionnaire are designed to support agencies to carefully consider their approach to procurement (including where algorithms may be provided free of charge). Further clarification of those questions is provided below.

When completing the AIA Questionnaire, please attach or link to any relevant evidence provided by the selected supplier(s) to support any claims about their responsible and ethical approach to algorithms and related data. That includes details of whether the supplier's governance and ethical positions align with New Zealand regulatory requirements (including the Privacy Act 2020), the Algorithm Charter and principles of open government.

As with the rest of the algorithm lifecycle, a multi-disciplinary approach to algorithm procurement is key to ensuring the selected algorithm is the best fit for the Purpose and key risks have been identified and addressed or accepted.

Key considerations, risks and mitigation options

Potential risks

- A lack of supplier algorithm or AI governance potentially resulting in unreliable algorithms that produce inaccurate or unfair outcomes
- A lack of supplier transparency on their data and how the algorithm works.
- 'Black box' algorithms where details of how the algorithm generates its outputs are unavailable, whether for commercial confidentiality or because they are so complex this is not well understood. This can lead to vendor lock-in and risks undermining the agency's ability to meet its 'Transparency' commitment under the Charter.
- Privacy issues arising from how and why the supplier collected personal information used in training data, as well risks arising from supplier access to agency-held data.

Nature of procurement (Question 9.2)

Please provide a clear description of exactly what you are procuring. For example, is the algorithm or application:

- custom-made by an external supplier for your agency
- a commercial off-the-shelf solution
- an open-source pre-trained model
- a mix of the above or something else (please specify).

CASE STUDY – School bus planning

To improve the optimisation of school bus route design, the Ministry of Education uses an algorithm to develop, standardise, automate, and maintain school bus routes.

Using licensed software, the algorithm calculates the most effective route for pick-up and drop-off of students, drawing on up-to-date information about road changes and speed limits. This has made bus travel more efficient for children and communities, led to significant efficiencies in planning time for bus routes and bus travel times, saved $20 million for taxpayers each year and reduced greenhouse gas outputs.

Source: Algorithm Assessment Report (Internal Affairs, Stats NZ), page 12

Evaluation criteria (Question 9.3)

Ensure you are using robust criteria to evaluate potential suppliers that address the specific risks associated with algorithms (and AI applications in particular). For example:

- Does the supplier have its own Responsible AI programme or similar that includes a comprehensive set of policies and procedures, like guidelines for ethical AI development, risk assessment frameworks and monitoring and auditing protocols that align with Charter commitments?
- Has the supplier provided any evidence of appropriate security and privacy controls?
- Has the supplier provided evidence of how they monitor, identify and manage bias risks?
- Is the operation of the algorithm or system clearly explainable?
- Is the supplier being transparent about their data collection practices and the nature and source of their training data?
- Will user data be used by the supplier to continue training the algorithm/model? If so, is it possible to opt out of this?
- What assurances can the supplier provide about not using exploitative labour practices for tasks such as data labelling and moderation?
- What evidence can they provide as to the safety and reliability of the algorithm?
- Can they demonstrate compliance with any relevant regulatory requirements or industry standards (for example, the NIST AI Risk Management Framework)? Has this been independently verified?

To drive the right behaviours by prospective suppliers of algorithms and AI tools, agency procurement teams should consider placing contingencies on supplier access to public sector procurement opportunities. For example, including requirements for supplier to demonstrate compliance with Charter commitments and ensure the explainability and interpretability of algorithms (for example, by sharing test results and explanations). This should be backed up with equivalent contractual obligations.

Contracts (Question 9.4)

As a general rule, you should aim to ensure your contracts with external suppliers include clear obligations, indemnities and liability positions to ensure appropriate allocation of risk between the parties.

You will need to work closely with your Procurement and Legal teams to achieve this. To the extent possible – and depending on the relative bargaining power of the parties - you should aim to incorporate the evaluation criteria outlined above into the supplier's obligations and require the following binding obligations from suppliers.

- Compliance with applicable laws and industry standards (and the supplier's own responsible or trustworthy AI principles, if any)
- Maintenance of high data quality and appropriate data security and privacy
- Responsibility for appropriate algorithm performance
- Responsibility for ensuring the algorithm is explainable and its outputs are reasonably comprehensible. You should aim to ensure commercial confidentiality and IP protection is not used as a barrier to transparency and explainability.
- Clarity in respect of the ownership of intellectual property rights and risk allocation for any third-party IP infringement risks (particularly for Generative AI applications)
- Provision of audit rights
- Appropriate indemnities and liability positions aligned with the extent of risk.

Supplier training data (Question 9.5)

If a supplier's algorithm is trained on data collected overseas that is not representative of the New Zealand populations to which it will be applied, this could lead to biased or discriminatory outputs. For example, a facial recognition model trained on images from North American or Chinese populations is likely to struggle to accurately identify New Zealand populations.

You should therefore aim to get as much visibility as possible of the supplier's training data sources so you can understand the potential for unfair outcomes and how those can be identified and addressed.

The reality is, however, that many suppliers will refuse to share this information for commercial confidentiality reasons. In addition, the relative bargaining power between the parties may be such that it is just not possible to get this information.

Where that is the case, you will need to consider the context and potential risks and clearly outline the issues in the AIA Questionnaire, including why the supplier refuses to be transparent about their training data. In some instances, this may not result in significant risk or there may be scope to focus additional efforts on identifying and mitigating downstream harms to compensate for this lack of visibility. The acceptability of the risk profile will need to be decided on a case by case basis by the AIA decision maker with input from the Project team.

Generative AI tools

Generative AI tools are able to generate high-quality content extremely quickly and efficiently. However, there are various specific issues and risks to be aware of.

- **Inaccuracy:** Generative AI tools like ChatGPT can instantly produce convincingly human-like written material. However, due to the way these tools work, the material they produce can often contain errors, be entirely fabricated and contain bias.
- **IP infringement:** Generative AI tools are trained on vast volumes of content, creating a risk of infringement of third party copyright.
- **Confidentiality and privacy:** If confidential or personal information is to be entered into a generative AI, you will need to ensure that such information is securely held and not accessible by the supplier or used for continued training of their model.
- **Worker exploitation:** AI tools may be fuelled by poorly paid workers in developing countries tasked with highly repetitive work such as labelling data or reviewing and flagging toxic content for moderation purposes. Such content may include murder, suicide, torture and child sexual abuse imagery, leading to mental health implications as a result of having to review disturbing material.

Where Generative AI tools are proposed, please refer to the following New Zealand guidance, which will be updated as the technology evolves.

- Guidance from the Office of the Privacy Commissioner on Generative Artificial Intelligence
- Joint System Leads tactical guidance on Generative AI

Agencies using third-party generative AI-related tools should consider how best to address these issues, seeking appropriate legal, privacy and other advice as necessary.

Open-source AI solutions

Open-source software is designed to be freely available to the public for use, modification and distribution by anyone. Open-access Generative AI systems enable anyone to develop their own apps for free. The benefits of open-source solutions include greater transparency, encouraging innovation through open collaboration and increasing adoption by reducing barriers to entry.

There are also obvious benefits in being able to access these solutions for free, making them particularly attractive in the public sector and to other organisations with fewer resources to develop or procure algorithms and AI systems themselves.

The Joint System Leads tactical guidance on Generative AI sets out various risks associated with using open-source or open access Generative AI solutions. It recommends exercising caution when using open-source AI, including taking steps to assess the testing, maintenance and governance of open-source AI software to ensure it is secure, appropriate, of high quality, and properly supported over time.

It also recommends ensuring you are following government procurement rules when sourcing Generative AI tools, conducting market research on suppliers and their offerings and including specific commercial protections in supplier contracts (including for privacy, security and ethical risks, technology obsolescence, vendor lock-in, and reliance on third-party provided services/AI).

Note that many of these issues are also relevant when procuring "closed-source" solutions, particularly in relation to vendor lock-in and reliance on third-party provided services.

Supplier access and use (Question 9.6)

Data and privacy breach risks are likely to increase if the supplier has access to your production data, so please detail why any such access may be necessary and what security measures and other relevant controls will be implemented to protect that data from unauthorised access, use or disclosure.

Who will be responsible for those security measures and what contractual obligations have been placed on the supplier in relation to such access?

Please also clarify whether the supplier plans to use or keep your production and/or algorithm user data (that is, any data input by the user of the algorithm, including where an algorithm is made directly available to the public to engage with). That includes where the supplier may wish to use that data to continue training its own algorithms/models, as may be the case with forms of Generative AI.

Suppliers should not be allowed to use the relevant data to train their proprietary models without a very robust justification and appropriate controls. Consider and describe the privacy, confidentiality and Māori Data Sovereignty risks of the proposed approach, what risk mitigants and controls are planned and how the issues have been addressed in the contract. Please ensure alignment with the Joint System Leads tactical guidance on Generative AI.

Please also consider and describe who will obtain the commercial benefits arising from the supplier's access to and use of agency data, including whether those benefits will be shared with the agency, Impacted People or other relevant stakeholders.

ADDITIONAL GUIDANCE

In addition to the *Government Procurement Rules*, you may find assistance from the following AI-focused international guidance.

- The World Economic Forum's *AI Procurement in a Box* is a practical guide for government agencies procuring AI tools that focuses on innovation, efficiency and ethics. The WEF argues this approach will not only accelerate the adoption of AI but also drive the development of ethical standards in AI development and deployment more generally. The guide includes Principles-based guidelines for AI procurement and a workbook for policy and procurement officials, include risk assessment criteria and case studies.
- The UK government's *Guidelines for AI Procurement* provide a further set of guidelines on how to buy AI technology as well as insights on tackling AI challenges that may arise during procurement. See also the UK government's *A guide to using artificial intelligence in the public sector*.

Algorithm performance, testing and monitoring

Why this is important

Appropriate testing, monitoring and ongoing review of algorithms is critical to ensuring appropriate performance in line with the defined Purpose and to minimise the risk of unfair outcomes. An algorithm may need to be re-trained if it is not producing the expected or desired outputs.

General guidance

Performance and testing (Question 9.7)

When completing the AIA Questionnaire, please attach or link to documentation detailing the technical features of each algorithm to facilitate understanding of how the algorithm arrives at its outputs.

Please also provide details of the results of any **testing** that has already been conducted and any that will occur going forward across the algorithm lifecycle.

- Appropriate performance ("this is what good looks like") should be clearly defined and documented based on the sensitivity and use of the algorithm and data in question. For example, when testing for unfair outcomes, you should first agree and document what "fairness" means in the context of the impacted groups, the Project's business goals and the overall Purpose for using the algorithm.
- Appropriate metrics to measure the algorithm's performance ("this is how you measure 'good'"), accuracy and unfair outcomes should also be defined and documented. For example, machine learning systems can be measured in numerous ways - see for example here and here.
- Quantitative test results addressing fairness may include evaluations of the algorithm's (or, as appropriate, the wider system's) accuracy for certain communities and demographic groups, such as women, Māori, Pasifika and socio-economically deprived groups. For example, has the algorithm been tested for differential accuracy or validity by subgroups (for example, ethnicity or gender)? Is there potential for a disproportionate benefit or disproportionate harms to one group or another in applying or interpreting the results? If so, how do you propose to mitigate this? Please include a specific Māori lens in your response.

Testing details should include an explanation of how similar the testing data and environment are to their real-life or production equivalents.

Key risks and mitigation options

Testing metrics (Question 9.7)

Some possible testing metrics that may be appropriate include the following. Note that accuracy might not be the most appropriate metric depending on the algorithm, in which case other metrics may be required.

1. ***Accuracy:*** The accuracy of an algorithm, also referred to as the 'error rate', is the proportion of examples for which it generates a correct or (in the case of an error

rate) an incorrect output. How accuracy is measured may be context dependent – in some instances, the choice of an acceptable error rate or accuracy level can be adjusted according to the use-case specific needs of the application.

Where machine learning models have been trained using supervised learning methods, please provide details of the algorithm's performance (percentage accuracy) on hold-out test data (that is, data that was not used during training, noting that this data should be reflective of the production data to be used). See the discussion on page 8 of MSD's *'Data Science Guide for Operations'* as part of the MDL for further discussion on using hold-out data to evaluate models.

Biometric recognition systems use "probabilistic matching", which is the process of using statistical analysis to determine the overall likelihood that two records match. This typically involves comparing input data (for example, in the context of facial recognition, a newly captured facial image) to the stored data (for example, a facial image recorded on a facial recognition watchlist).

Unlike typical binary decision-making systems used for traditional verification methods such as for passwords (see discussion below), biometric recognition systems like facial recognition involve a range of factors such as lighting differences, image quality and camera angle that can influence the accuracy of a match.

To reflect those potential accuracy issues, confidence scores are often used in biometric technologies to indicate the likely accuracy of the output. A confidence score tells you how confident the underlying algorithm is that it has extracted the correct value. It is typically provided as a percentage, with a higher score representing greater confidence. For example, image matching tools that use computer vision to compare images and indicate where there is a match will often use a confidence score to indicate the expected accuracy of the match – for example, a confidence score of 99% indicates a higher likelihood of a genuine match than a score of 65% would.

In those circumstances, there is often the ability to set decision thresholds that determine at what level the system will suggest a match – for example, a match will only be suggested for confidence scores over 85% because a lower score would unreasonably increase the risk of "false positive" errors. False positive errors occur when a system incorrectly observes a case as positive when it shouldn't – that is, a match is suggested but the image does not match the person. For example, there have been numerous recorded instances of

CASE STUDY – Facial recognition false positive results in arrest of pregnant woman

An eight-month pregnant woman was wrongfully arrested in Detroit for carjacking and robbery after a facial recognition system falsely identified her as the attacker.

The facial recognition system matched surveillance footage from a petrol station to the woman's mugshot from a 2015 arrest for driving with an expired licence. A human analyst confirmed the system's suggested match, as did the victim - indicating potential automation bias.

At the time of reporting, the woman had filed a lawsuit for wrongful arrest and imprisonment.

Source: https://www.usatoday.com/story/news/nation/2023/08/08/facial-recognition-technology-wrongful-arrest-pregnant-woman/70551497007/

facial recognition systems used in the US incorrectly identifying black people.

False negative errors occur when a system incorrectly observes a case as negative when it should be positive (that is, the image does match the person but the system did not recognise them)

False positives can have significant impacts, particularly in the context of facial recognition technology, and may be more likely to arise where training data does not come from New Zealand and/or where systems have not been tested on New Zealand faces. See a relevant discussion in this article.

Before deploying an algorithm or system that uses a confidence score, such as a biometric system, you should ensure you understand the potential implications of false positives and false negatives. You should also consider the possible impact on those who may use, rely on or be affected by the system.

Please provide details of how any decision thresholds are determined, the percentage of false positives and false negatives on held-out test data and any associated risks. Please detail any trade-offs between false positives and negatives and the extent to which such trade-offs could be considered reasonable in the context of the Purpose.

2. ***Binary decisions:*** Some algorithms are trained to make a binary judgement (that is, to categorise data into one of two groups based on certain criteria). For example, traditional password verification methods compare an input value (what you type) with a stored value (your password). If the input exactly matches the stored value, then access is granted.

   For binary algorithms, please provide details of:

   - the percentage of false positives and false negatives on held-out test data; and
   - the F-score for these systems (that is, a formula that combines precision and recall to measure an algorithm's accuracy on a dataset) with an explanation of what this means.

It's important to note the comments in MSD's *'Data Science Guide for Operations'* that each type of predictive error has different implications in practice. Some measures of accuracy treat different types of errors – such as false positives and false negatives - as if they have the same importance. However, in practice, different errors will tend to have different real-world implications.

MSD gives the example of an analytical model for benefit fraud that flags potential cases of suspicion. While a false negative would lead to non-detection of an actual benefit fraud, a false positive could lead to incorrect conclusions regarding a beneficiary, which could have significant implications for the beneficiary (recall the Dutch benefit case study earlier and the Robodebt case study below).

Depending on the Purpose, business context and potential harms, there might therefore be a greater need to reduce false positives at the expense of compromising on the false negative error rates or vice versa. The full range of potential outcomes, including unfair outcomes, needs to be identified and considered by a multi-disciplinary group in the context of the Purpose when evaluating the accuracy of a model. The issues and decision should be documented. Data scientists can then give whatever weighting the group has determined is most appropriate to the different types of errors.

CASE STUDY - Robodebt

The “Robodebt” scandal occurred after Centrelink, the Australian government department responsible for delivering social security payments and services, used an algorithm to identify discrepancies between income declared to the Australian Taxation Office and reported to Centrelink. Debt notices were automatically generated in relation to any discrepancies.

In 2021 a Federal Court Judge approved a settlement of A$1.8 billion relating to nearly half a million false accusations of benefit fraud. Much like in the Netherlands, the human impacts of Robodebt were significant - many victims experienced mental health impacts and there were several suicides.

A Royal Commission into the Rododebt Scheme made 57 recommendations, including new legislation aimed at ensuring that public service algorithms and automated decision systems are fit for purpose, lawful, fair, and do not adversely affect human and legal rights; establishment of a body to monitor and audit automated decision-making; and requirements for transparency and to make business rules and algorithms available to independent expert scrutiny.

The scandal severely damaged the reputation of Centrelink and eroded public trust in the government’s ability to manage social services. The Royal Commission report states that:

> *Robodebt was a crude and cruel mechanism, neither fair nor legal, and it made many people feel like criminals. In essence, people were traumatised on the off chance they might owe money. It was a costly failure of public administration, in both human and economic terms.*

Source: https://www.theguardian.com/australia-news/2023/mar/11/robodebt-five-years-of-lies-mistakes-and-failures-that-caused-a-18bn-scandal

Ongoing monitoring (Question 9.8)

Full monitoring across an algorithm’s lifecycle is critical to enabling the identification of performance issues. Algorithmic outputs should be regularly tested to ensure the performance that was established and confirmed during development is maintained, despite any changes in the algorithm’s operational environment. Bias and other unfair outcomes may only become apparent after an algorithm is in operation.

For any continuously deployed predictive algorithm, protocols should be in place for regular re-evaluation and the regular gathering of new training data to keep the system up to date (Gavaghan, C., Knott, A., Liddicoat, J., Maclaurin, J., & Zerilli, J. (2019) Government Use of Artificial Intelligence in New Zealand).

Appropriate documentation should also be maintained to record:

- assessment metrics and methodology
- an intervention plan in case performance issues or biased outputs are identified

- how the monitoring and intervention will be implemented when the algorithm is deployed
- responsibility in case of failure, particularly where external suppliers are involved. Appropriate accountability should be addressed in contractual arrangements with such suppliers.
- process or log for monitoring and capturing user complaints and comments
- evidence of successful tests against benchmarks.

Algorithms may need to be retrained to guard against issues like concept drift, where the target variable that an AI model is trying to predict changes over time in unforeseen ways, making the predictions less accurate over time.

Decisions on how regularly algorithms are retrained should be made by developers and data scientists in collaboration with the business owner and your privacy, legal and ethical advisers to ensure an appropriate breadth of input. Consider alignment with the defined Purpose and whether data remains current and representative.

It is important to ensure you have adequate resources throughout the algorithm's lifecycle to properly maintain the algorithm, conduct monitoring activities **and** to resolve any issues identified as a result of such monitoring.

Please also confirm the life expectancy of the algorithm and when it, or the data that powers it, is likely to become obsolete or liable to cause harm such that it will need to be retired or replaced.

ADDITIONAL GUIDANCE

For more detail on algorithm selection and optimisation, see the Ministry of Social Development's *Data Science Guide for Operations* that forms part of its Model Development Lifecycle.

10. Safety, security and reliability

Remember to attach or link to a copy of your Security Risk Assessment as well as answering the questions in the AIA.

Why this is important

To maintain trust and confidence in their use, algorithms need to be safe to use, secure, dependable and resilient in the face of change.

This is another area that is highly context dependent. The interlinked safety considerations of accuracy (discussed above in the *Data* section), reliability, security, and robustness ('safety') in your Project will depend on a variety of factors, including what algorithm and, if applicable, machine learning techniques will be used, how those techniques will be deployed, the nature and source of your data, how you have defined your Purpose and the problem you are trying to solve.

General guidance (Question 10.1)

Reliability: Reliability is a measure of consistency and can establish confidence in the safety of a system based upon the dependability with which it performs as intended, even with new data on which it has not been trained or tested previously.

Security: A secure system is capable of maintaining the integrity of the information within it. This includes protecting its architecture from unauthorised modification or damage to any of its component parts. A secure system remains continuously functional and accessible to its authorised users and keeps confidential and private information secure even under hostile or adversarial conditions.

Robustness: A resilient system maintains its functionality and performs accurately in a variety of environments and circumstances, even when faced with changed inputs or an adversarial attack.

The wide range of uses of many generative AI systems means that safety risks may need to be assessed more broadly than those of systems with more specific uses.

Key considerations and risks (Questions 10.2, 10.3)

Algorithmic and AI systems may be vulnerable to a range of threats, including the following.

Security flaws: You should avoid providing potentially insecure external parties with access to training or production data given data's critical role in powering algorithms. Internal or external parties could gain access to data, algorithms and their outputs and manipulate them to introduce deliberately flawed outcomes. Generative AI in particular can enable those with little or no coding experience to easily write functional malware. Third party browser plug-ins or extensions may inadvertently expose network environments.

Data or model poisoning: Adding inaccurate or misleading data to the training dataset or injecting undetectable defects into the algorithm to trigger incorrect outputs.

Adversarial attacks: Manipulation of an AI system to cause unreliable outputs. That includes “prompt injection attacks” in the context of Generative AI, where an attacker hijacks and controls a language model’s output and can gain access to confidential and personal information.

Misinformation and disinformation: A particular risk in the context of generative AI, this may occur where content is not clearly identified as being AI-generated and could potentially lead to confusion (‘misinformation’) or deception (‘disinformation’). Threat actors can use this in scams or fraudulent campaigns against individuals and organisations.

As this is a highly technical area, you should consult with your Security team to ensure that safety risks have been taken into account and mitigated throughout the algorithm lifecycle.

Risk mitigation options (Questions 10.3, 10.4)

All algorithm projects should be designed with safety and security in mind and security advisers should be included in workshops and other discussions forming part of the AIA process.

1. **Identify** and **define**:
   - the algorithm’s key vulnerabilities and risks (for example design faults, technical faults, cyber-attacks)
   - possible consequences
   - risk metrics and risk levels for each specific use case.

2. **Implement** ways to continuously measure and assess risks and potential attacks over the algorithm’s lifecycle. That includes:
   - testing, validating, verifying and monitoring the safety of the algorithm on an ongoing basis
   - protecting data with appropriate access restrictions and encryption both in transit and at rest
   - considering targeted risk mitigation strategies for AI tools, including model hardening, runtime detection, hard-wiring mechanisms into the system that enable human override and system shut-down, and continuous inspection and monitoring
   - conducting penetration testing
   - managing potential external supplier risks, including conducting appropriate due diligence (as discussed above in relation to procurement) to ensure external party components are adequately verified and appropriate security protections are in place
   - training staff on the key risks and how to approach them
   - performing safety self-assessments to evaluate how a Project’s design and implementation practices line up with the safety objectives of accuracy, reliability, security, and robustness. These self-assessments should be recorded to facilitate review and re-assessment.

3. **Build and test a response plan**. This includes clear roles and responsibilities, processes, and procedures to address the risks. Where an incident does occur you

should have a communications plan in place to alert the public and ensure as transparent an approach as possible.

- Consider running high-risk algorithms and AI applications in a ‘sandbox’ or other safe environment before full deployment to ensure they are working as anticipated.
- For business continuity purposes, it may be sensible to maintain alternative versions of the algorithm that can be put into operation if the principal version has to be taken offline for any reason and critical operations will be affected. Manual or other methods of accomplishing the task should also be available as a backup.

ADDITIONAL GUIDANCE

Resources for combatting adversarial attacks are available at

https://github.com/IBM/adversarialrobustness-toolbox*

*Note: This link requires a login to Github

11. Community engagement

Why this is important

The Charter commits signatories to focus on ‘People’ by identifying and actively engaging with people, communities and groups who have an interest in algorithms and consulting with those impacted by their use.

The inclusive development of algorithms and consideration of a diverse range of perspectives - that agencies may not otherwise have access to - is likely to improve algorithmic performance, reduce many of the potential harms signposted in the AIA process, increase transparency and ultimately help build trust and confidence in government use of algorithms.

Key considerations, risks and mitigation options (Questions 11.1 to 11.3)

As with many other aspects of this assessment, the extent of community engagement required for any given algorithm will depend on the context and Impacted People involved. The greater the impact, the more extensive any engagement should be.

You should consider how individuals and communities are likely to react to the use of the algorithm and whether you have sufficient ‘social licence’ to proceed with the Project. ‘Social licence’ includes being transparent about how data is being used and people trusting that their data will be used as they have agreed and accepting that enough value will be created.

The Data Futures Partnership suggests that organisations use their own judgement to decide whether active engagement is needed to achieve trust. However, the need to engage should be carefully considered where a planned use of data in an algorithm will:

- be a novel use for the community it will affect
- have a substantial impact on whānau, hapū, iwi, Māori communities or Pasifika
- have a disproportionate impact on people from small communities or people identifying as disadvantaged
- have an impact on vulnerable groups such as minors
- have the potential to have a serious impact on people's lives (for example, decisions about access to social housing or mortgages)
- involve sensitive information or
- be proposed by an organisation starting from a low level of trust (for example, following a serious breach of data security).

Effective public engagement will require identification of which kinds of diverse expertise are required and how those perspectives will be obtained and factored into the algorithm design. Agencies seeking to use high impact algorithms should seek input from as broad a range of stakeholders as possible – consider engaging with community and civil society groups, cultural representatives, academics, the private sector, the Government Chief Privacy Officer, the Office of the Privacy Commissioner, the Data Ethics Advisory Group, the Interim Centre for Data Ethics and Innovation and relevant groups and organisations.

Obtaining input from diverse communities about their own experiences will help ensure the algorithm is responsive to their needs. The people most impacted by an algorithm often have the least power but the best understanding of how to address potential risks.

If no consultation is planned, please explain why and what other methods are being adopted to ensure sufficiently diverse community perspectives are incorporated into the Project.

Consider what evidence you have that the use of the algorithm for the Purpose will be acceptable to Impacted People. If social licence is lacking, how is it proposed this will be developed?

ADDITIONAL GUIDANCE

- For further guidance and examples of algorithmic transparency reports, see the UK Centre for Data Ethics and Innovation's Algorithmic Transparency Recording Standard Hub.
- The DPUP provides helpful guidance and questions around consultation and diversity as part of the He Tangata and Manaakitanga principles.
  - **He Tangata** focuses on improving people's lives and ensuring that everything we do with data should contribute to the wellbeing of the individual or the community.
  - **Manaakitanga** is about respecting and upholding the mana and dignity of those who share their data, encouraging the incorporation of diverse interests, perspectives and needs and the inclusion and involvement of service users whenever possible.
- The Data Futures Partnership's *A Path to Social Licence – Guidelines for Trusted Data use*

12. Transparency and explainability

Why this is important

The Charter requires signatories to maintain transparency by clearly explaining how their decisions are informed by algorithms.

It provides that signatories may maintain transparency by providing plain language documentation about the algorithm, making information about the data and processes available and publishing information about the collection, security and storage of data.

Transparency and explainability are important because they enable public scrutiny and greater accountability of public sector decision-making processes using algorithms. This enables those impacted by an algorithm to understand how and why an algorithm produces the outputs it does, better equipping Impacted People to challenge those outputs where necessary, including any decisions they consider to be unfair, unreasonable, inaccurate or unlawful.

This is particularly important for administrative decisions, where procedural fairness and satisfaction of due process are fundamental requirements. It also helps courts determine if an error of law has occurred and promotes increased public trust and confidence in the administrative process and use of algorithms.

Transparency around algorithm use will also help to maintain public trust and confidence. Best practice includes publication of completed AIA Reports to support transparency obligations under the Charter.

General guidance

'Transparency' is used in the Charter as an umbrella term encompassing the following concepts.

- **Explainability:** the ability to describe an algorithm's decision-making process in a way that is understandable to humans. Since explainability is highly contextual - a data scientist is likely to have a different understanding to that of a consumer - there may be no need in certain circumstances for explainability (for example, optimising a data warehouse) whereas in other cases it may be crucial (for example, explaining to a taxpayer why they were charged a penalty).
- **Interpretability**: the ability to understand the mechanics of an algorithm's decision-making process. A fully interpretable model is one whose decisions, given a known set of inputs, can be reproduced by a human using tools such as a spreadsheet or even pen and paper (often referred to as "local interpretability"). "Global interpretability" refers to the ability to describe the decision-making process of the entire model in a human-readable format, such as a decision tree, mathematical equation, or block of code.
- **Transparency/disclosure:** communicating about the use of an algorithm and the role it plays in any decision-making process.

AI models in particular are often described as "black boxes", meaning they are very difficult to practically interpret, explain and understand. This can be either due to their complexity (such as deep neural networks like Large Language Models (LLMs) or as a result of the "closed source" nature of the model in question.

CASE STUDY - UK A-level exams*

After UK students were unable to sit their A-level university entrance exams due to COVID lockdowns, an algorithm was used to determine student marks. It aimed to achieve a similar distribution of marks to previous years.

The calculations for each student included not only each student's but also their school's past performance – favouring those from wealthier areas and private schools.

The grades of nearly 40% of students were reduced, leading to accusations that the algorithm was opaque, unfair and discriminated against students from disadvantaged backgrounds. The UK government abandoned use of the algorithm after protests and intense criticism of its approach.

Source: https://www.nytimes.com/2020/08/20/world/europe/uk-england-grading-algorithm.html

*Note: link to this case study requires a subscription to the New York Times

"Glassbox" models are machine learning models where the underlying mechanisms for generating predictions and the reasoning behind the predictions made by those models can be easily and completely explained in a way a human can understand.

Key considerations and risks

The AIA Questionnaire asks you to consider how information about the algorithm will be made available to both Impacted People and the wider public. If it is not possible to readily communicate such information, then please explain why.

Where an algorithm is part of a wider system, please detail the extent to which it is possible to clearly identify which algorithm led to specific decisions or recommendations. If this will not be possible, or at least not easily, then please explain why.

While the Official Information Act 1982 provides a legal right of access to the reasons for decisions by official agencies (section 23), and the Privacy Act 2020 enables individuals to request access to and correction of their personal information, both of those transparency mechanisms are only triggered upon request by the relevant individual. While they are important to consider, the focus of the Charter and this AIA process is on *pro-active* transparency.

Privacy obligations (Question 12.1)

Where the algorithm uses personal information, you will have transparency obligations under the Privacy Act 2020.

Agencies are required to take reasonable steps to ensure individuals are aware that information about them is being collected and what it's being used for. The Guidance from the Office of the Privacy Commissioner on Generative Artificial Intelligence states:

> "If the generative AI tool will be used in a way likely to impact customers and clients and their personal information, they must be told how, when, and why the generative AI tool is being used and how potential privacy risks are being addressed. This must be explained in plain language so that people understand the potential impacts for them before any information is collected from them. Particular care must be taken with children."

Commercial sensitivity objections (Question 12.1)

It will be easier to ensure transparency with algorithms that have been developed in-house. Many external suppliers may refuse to disclose the inner workings of their algorithms and related systems on the grounds of commercial sensitivity. Intellectual property rights may also be used to prohibit access to proprietary code and training data.

As much as possible, you should aim to ensure contracts for externally procured algorithms require suppliers to facilitate transparency and explainability by providing access to this information so agencies can meet their Charter obligations.

Complex neural networks (Question 12.2)

Some forms of AI, such as neural networks, present particular challenges from a transparency and explainability perspective because it can be unclear precisely how or why a particular output has been generated.

For example, LLMs like ChatGPT use neural networks to identify the patterns and structures in existing data to generate new and original content. But due to the size of the models, the fact that the training data is kept confidential and the closed-source nature of their implementation, it is very challenging to understand how and why they produce the outputs they do.

Explainability is still an active field of research. If there are legal requirements for algorithm-based decisions to be explainable and that is not possible with the proposed algorithm, then an alternative algorithm or approach may be necessary.

Generative AI (Question 12.1, 12.2)

The Joint Systems Leads tactical guidance on Generative AI recommends agencies are open and transparent about how and why Generative AI is being used, ensuring processes are in place to respond to citizen requests to access/correct information.

It notes that the Public Service is often held to a higher standard than the privacy sector so agencies should consider how to assure transparency, accountability, and fairness in how they are using and applying Generative AI, whether directly or as part of a wider technology solution.

The Guidance from the Office of the Privacy Commissioner on Generative Artificial Intelligence states:

> "If the generative AI tool will be used in a way likely to impact customers and clients and their personal information, they must be told how, when, and why the generative AI tool is being used and how potential privacy risks are being addressed. This must be explained in plain language so that people understand the potential impacts for them before any information is collected from them. Particular care must be taken with children."

Risk mitigation options

The extent of transparency and explainability required for any given algorithm is context dependent, requiring consideration of the nature of the algorithm and situation in which it's being applied, the types of data inputs (particularly where personal information is involved), the types of data outputs, who it impacts and the nature of those impacts.

As a general rule, you should aim to communicate the following high-level information.

- **The fact an algorithm is being used and why**. Give a basic overview of the purpose of the algorithm, including:
- What issue or problem you're aiming to solve and how the algorithm assists this
- The justification or rationale for using the algorithm.
- How the algorithm is being used, including explaining:
- how it works
- how it is used and by whom
- the nature of its outputs (for example decisions, predictions) and how and why they are produced, including the logic or reasons used to generate the outputs.
- **Data:** What data is used in relation to the algorithm, where it comes from and how it will be handled, secured and stored.
- **Responsibility:** Who is involved in the development, management and implementation of the algorithm, the extent of human oversight and the relevant human point of contact. That includes information about how people can find out more about the algorithm or ask a question.

- **Impact:** Who the algorithm will impact and how as well as how those impacts are being monitored.
- **Accuracy, safety and fairness:** Steps taken across the design and implementation of the algorithm to maximise the accuracy and reliability of its outputs and to ensure those outputs are fair and unbiased.

You might like to consider adapting algorithm or AI "Nutrition Face Labels" for your particular context. These are intended to provide a more transparent view of how such systems use people's personal information to produce the results they do. See here for more information.

ADDITIONAL GUIDANCE

- See the AI Forum NZ's White paper on Explainable AI – building trust through understanding
- For further guidance and examples of algorithmic transparency reports, see the UK Centre for Data Ethics and Innovation's Algorithmic Transparency Recording Standard Hub.
- You may also want to refer to the *Mana Whakahaere* principle in the DPUP, which supports the 'Transparency' Charter commitment. That principle focuses on empowering people by giving them choice and enabling their access to and use of their data and information. The webpage on the Mana Whakahaere principle provides more detail on how to achieve this.
- The UK's Information Commissioner's Office and The Alan Turing Institute have provided helpful guidance on Explaining decisions made with AI, which aims to give organisations practical advice to help explain the processes, services and decisions delivered or assisted by AI to the individual affected by them. Although it is based around English data protection legislation, you may still find the guidance helpful.

Glossary

Accountability: The requirement for organisations and their leaders to be responsible for their actions and decisions, including explaining and justifying their conduct. When it comes to algorithms, this includes ensuring that those that build, procure and use algorithms:

- can justify and are ultimately answerable for such usage and their impacts
- ensure algorithmic systems operate in a manner that is ethical, fair, transparent and compliant with applicable rules and regulations
- can face consequences for such use.

Adversarial Attack: Adversarial attacks on machine learning models maliciously modify input data to provoke a misclassification or incorrect prediction. For example, by undetectably altering a few pixels on a picture, an adversarial attacker can mislead a model into generating an incorrect output (like identifying a panda as a gibbon, or a 'stop' sign as a 'speed limit' sign) with an extremely high confidence. While a good amount of attention has been paid to the risks that adversarial attacks pose in deep learning applications like computer vision, these kinds of perturbations are also effective across a vast range of machine learning techniques and uses such as spam filtering and malware detection.

Agency: A generic term used to refer to New Zealand government entities across the public sector.

AIA/Algorithm Impact Assessment: The goal of an AIA is to mitigate potential harmful impacts of an algorithm initiative or deployment, recognising any potential risks and addressing them before implementation. AIAs are intended to enable public agencies to better understand, categorise and respond to the potential harms or risks posed by the use of algorithms prior to their use.

AI/Artificial Intelligence: A broad term used to describe an engineered system where machines learn from experience, adjusting to new inputs and potentially performing human-like tasks. The term is used in the AIA documentation as an umbrella term for a range of technologies and techniques that involve programming computer software to execute algorithms that can recognise patterns, reach conclusions, make informed judgments, optimise practices, predict future behaviour, and automate repetitive functions. This includes machine learning, natural language processing, generative AI (for example ChatGPT), computer vision and biometric technologies.

AI model: An AI model is a program or algorithm that has learnt from a set of data to recognise certain types of patterns. This allows it to reach a conclusion or make a prediction when provided with sufficient similar information.

AI system: Any AI-based component, software or hardware, often embedded as components of larger systems. For the purposes of this report, we refer to an AI system as a sociotechnical system, which may be made up of one or several algorithms. AI systems may use automated reasoning to aid, replace or augment human decision-making.

Algorithm: A procedure or formula for solving a problem or carrying out a task. Although they can be used in a non-digital context, the AIA documents use this term to describe a computational procedure or set of instructions and rules designed to process data inputs and return an output, perform a specific task, solve a particular problem, or produce a machine learning or other AI model.

Below are some examples of how different agencies have defined 'algorithm' for their own purposes:

> ***Ministry of Business, Innovation and Employment (MBIE):*** Algorithms are the automatic decision-making processes used by computer programs to identify patterns in data, in order to assess alignment to a set of criteria or predict outcomes.
>
> ***NZ Police:*** An objective system in which data is taken in, converted into a different form and returned as a set of outputs, a score or a suggested decision.

Algorithm lifecycle: From business need and inception stage through design, development, testing, verification and validation to deployment, operations and retirement.

Algorithmic system: A system that uses one or more algorithms to produce outputs that can be used for making decisions.

Algorithmic tool: A product, application, or device that supports or solves a specific problem by using complex algorithms. The AIA process uses this as a deliberately broad term covering different applications of AI and complex algorithms.

Artificial General Intelligence/AGI: AI that is considered to have human-level intelligence and strong generalisation capability to be able to achieve goals and carry out a variety of tasks in different contexts and environments. AGI is still considered a theoretical field of research and contrasted with 'narrow' AI, which is used for specific tasks or problems.

Automated Decision-Making/ADM: Refers to the application of automated systems in any part of the decision-making process to replace the judgement of human decision-makers. Automated decision-making systems range from traditional non-technological rules-based systems to specialised technological systems that use automated tools to predict and deliberate.

Automation bias: errors people tend to make in highly automated decision-making contexts, when decisions are handled by algorithms and other automated aids and the human actor is largely present to monitor on-going tasks. The unquestioning acceptance of such decisions or recommendations can lead to system errors being overlooked, potentially leading to harm.

Bias: Systematic differences in treatment of certain objects, people or groups in comparison to others. In a social context, bias can be one of the main causes of discrimination and injustice.

There are many different types of bias, including the following non-exhaustive list (SA TR ISO/IEC 24027:2022 Information Technology – Artificial Intelligence – Bias in AI systems and AI aided decision making).

- ***Human cognitive bias***: Occurs when humans are processing and interpreting information and includes:
    - **Automation bias:** The tendency for humans to favour suggestions from automated decision-making systems and to ignore contradictory information made without automation, even if that information is correct.

  - **Confirmation bias:** A type of cognitive bias that favours predictions of AI systems that confirm pre-existing beliefs or hypotheses.
  - **Societal bias:** Occurs when similar cognitive bias is held by many individuals in society. It can manifest when machine learning models learn or amplify pre-existing, historical patterns of bias in datasets or when cultural assumptions about data are applied without regard to cross-cultural variation.
  - **Systemic bias:** A form of societal bias embedded in systems such as society, a particular culture, or an organisation.
- ***Data bias*****:** Data properties that, if unaddressed, can lead to AI systems that perform better or worse for different groups. It arises from technical design decisions and can be caused by human cognitive bias and training methodology. Like human cognitive bias, there are numerous types of data bias such as:
  - statistical bias (includes selection, sampling, coverage and non-response bias)
  - data label and labelling process
  - non-representative sampling
  - missing feature of labels
- ***Computational bias*****:** a systematic error or deviation from the true value of a prediction that originates from a model's assumptions or the data itself.

Black box: An AI system where the data entered is known, and the decisions made from that data are known, but the way in which the data was used to make the decisions is not understood by humans.

Charter: The Algorithm Charter for Aotearoa New Zealand, a commitment to ensuring New Zealanders have confidence in how government agencies use algorithms.

Computer vision: A field of AI that enables computers to process and analyse images, videos and other visual inputs. For example, facial recognition technology.

Confidence score: The use of AI often involves estimation, such as the probability that the output is a correct answer to the given input. Confidence scores are a way of quantifying the uncertainty of such an estimate. A low confidence score associated with the output of an AI system means that the system is not too sure that the specific output is correct.

Data: A type of information (especially facts or numbers) that is collected to be categorised, analysed, and used to help decision-making.

Data subject: In the context of personal information, the individuals that the personal information relates to.

Discrimination: Unequal treatment of a person based on belonging to a category rather than on individual merit. Discrimination can be a result of societal, institutional and implicitly held individual biases or attitudes that get captured in processes across the algorithm lifecycle or represented in the data underlying algorithmic systems. Discrimination biases can also emerge due to technical limitations in hardware or software or in the very context in which the AI system is used. As many forms of biases are systemic and implicit, they are not easily controlled or mitigated and require specific governance and other similar approaches.

Disinformation: false or inaccurate information that is deliberately and often covertly spread to mislead or deceive and/or to influence public opinion.

Evaluation criteria: As outlined in the Government Procurement Rules, the criteria used to evaluate supplier responses, including measures to assess the extent to which competing responses meet requirements and expectations.

Explainability: The ability to describe or provide sufficient information about how an AI system generates a specific output or arrives at a decision in a specific context.

Facial recognition: Facial recognition algorithms detect and analyse faces and create unique digital biometric templates that can be used to match and identify individuals in photos, videos and real time.

Fairness: Fairness is about more than the absence of bias. Fair decisions need to also be non-arbitrary, reasonable, consider equality implications, and respect the circumstances and personal agency of the individuals concerned.

False negative: A negative outcome that an AI model predicted incorrectly. For example, a failure to identify a previously enrolled individual in a First Response Time (FRT) system.

False positive: A result indicating a certain condition is present when in fact it is not. For example, wrongly identifying another person as the enrolled individual when using FRT.

Generative AI / GenAI: uses input data or user prompts and questions to generate material that closely resembles human-created content, such as written text, code, images, music, simulations and videos. Generative AI models work by analysing large volumes of training data to detect and replicate patterns and relationships in that training data so they can then match user prompts to the identified patterns and probabilistically "fill in the blank" by predicting and generating the next word in a sentence, feature of an image, and so on. ChatGPT is the most well-known, free, example of Generative AI.

Harm: Adverse consequences for people of an algorithm's deployment and operation in the real world. The AIA process aims to anticipate, identify and avoid potential harms.

Held-out test data: Data used to test an algorithm, which was not used during training.

Impacted People: Those individuals, groups and communities who are likely to be either directly or indirectly impacted by use of the algorithm.

Large Language Model/LLM: A type of generative AI system that uses deep-learning algorithms trained on very large textual datasets to generate human-like text. For example, OpenAI's ChatGPT or Google's Bard. LLMs are probabilistic in nature and operate by generating likely outputs based on patterns they have observed in the training data.

Machine learning: A sub-field of AI involving algorithms that enable computer systems to iteratively learn from and then make decisions, inferences or predications based on data. These algorithms build a model from training data to perform a specific task on new data without being explicitly programmed to do so.

Māori: The term Māori, as used in this user guide, includes all individuals and collectives self-identified or recognised as Māori, including hapū, iwi and hapori. (This definition is from Te Kāhui Raraunga's Māori Data Governance Model).

Māori data: Māori data refers broadly to digital or digitisable data, information or knowledge (including mātauranga Māori) that is about, from or connected to Māori. It includes data about population, place, culture and environment. Māori view data as a living tāonga (treasure) with immense strategic value. (This definition is from Te Kāhui Raraunga's Māori Data Governance Model).

Māori data governance: The principles, structures, accountability mechanisms, legal instruments and policies through which Māori exercise control over Māori data. (This definition is from Te Kāhui Raraunga's Māori Data Governance Model).

Māori data sovereignty: The inherent rights and interests that Māori have in relation to the collection, ownership and application of Māori data. (This definition is from Te Kāhui Raraunga's Māori Data Governance Model).

Material impact: In the context of the AIA process, this means something that could reasonably be expected to affect the rights, opportunities or access to critical resources or services of individuals, communities or other groups in a real and potentially negative or harmful way, or that could similarly influence a decision-making process with public effect. For example, decisions about the administration of justice or democratic processes that impact people, or decisions impacting people's access to education, social welfare, health, housing, ACC or immigration services.

Misinformation: false or inaccurate information, regardless of any intent to mislead or deceive.

Model: A model is an expression of an algorithm that represents what has already been learned from data by the algorithm. A machine learning model is a predictive algorithm whose exact nature has been learnt through training on input data.

Model hardening: Advanced techniques to combat adversarial attacks by strengthening the architectural components of the systems. This may include adversarial training, where training data is methodically enlarged to include adversarial examples, architectural modification, regularisation, and data pre-processing manipulation.

Natural language processing: A sub-field of AI that helps computers understand, interpret and manipulate human language by transforming information into content. It enables machines to read text or spoken language, interpret its meaning, measure sentiment and determine which parts are important for understanding.

Personal information: Information about an identifiable individual. It covers both information that is simply about a person (for example, eye colour) and information that may also identify them (for example, their name). The information does not need to name the individual, as long as they are identifiable in other ways, like through their home address.

Privacy Act: The Privacy Act 2020 provides the rules in New Zealand for protecting personal information and puts responsibilities on agencies and organizations about how they must do that. For example, people have a right to know what information your agency holds about them and a right to ask you to correct it if they think it is wrong. For more information, see the website of the Office of the Privacy Commissioner and the Privacy Act 2020.

Privacy Impact Assessment (PIA): Similar to an AIA, a privacy impact assessment or PIA is a tool used by agencies to help them identify and assess the privacy risks arising from their collection, use or handling of personal information. A PIA will also propose ways to mitigate or minimise these risks.

Production data: Data that is used or produced during the normal day-to-day operations of the agency.

Prohibited grounds: Prohibited grounds for discrimination under the Human Rights Act 1993, including discrimination on the grounds of sex, marital status, religious or ethical belief, colour, race, ethnic or national origin, disability, age, political opinion, employment or family status, and sexual orientation.

Project: The project described in the AIA Questionnaire.

Proxy variable: A variable that is not in itself directly relevant but which is used instead of a variable that cannot be measured or is difficult to measure. Proxy variables may represent, or correlate with, other variables, potentially including sensitive attributes or prohibited grounds. For example, a postcode could operate as a proxy for race or height and weight as proxies for gender with potential for bias or discrimination.

Purpose: In the context of the algorithm, this refers to how and why the algorithm helps achieve the objectives of the Project in the relevant business context. This is particularly important when considering accuracy, potential biases and other unfair outcomes, which tend to be highly contextual.

Reinforcement learning: The process of training a model by using trial and error, where the system receives rewards for performing well and punishments for performing poorly.

Reliability: Where an algorithm or AI system performs its intended function consistently, accurately and as expected, even with new data on which it has not been trained or tested previously.

Risk: The composite measure of the probability of an event occurring and the magnitude of its consequences. In the context of algorithms and AI, ‘risk’ is often used to refer to the risks to an agency or organisation, such as compliance, legal, reputation or financial risk. ‘Harm’ is more typically used to refer to the negative impacts on Impacted People arising from the use of an algorithm or AI application.

Robustness: A resilient system that maintains its functionality and performs accurately in a variety of environments and circumstances, even when faced with changed inputs or an adversarial attack.

Semi-supervised learning: The process of training a model where the training data is made up of both labelled and unlabelled data. Semi-supervised learning is often done by manually labelling a relatively small part of a large unlabelled dataset.

Sensitive personal information: Sensitive personal information is information about an individual that has some real significance to them, is revealing of them, or generally relates to matters that an individual might wish to keep private. The Privacy Act does not prescribe fixed categories of “sensitive” personal information and any personal information can be sensitive depending on the particular context and surrounding circumstances, including cultural perspectives. Certain types of personal information are inherently sensitive however, including health, genetic, biometric and financial information. The personal information of children and young people is also sensitive, given their inherent vulnerability and more limited agency than adults.

Social licence: When people trust that their data will be used as they have agreed, and accept that enough value will be created, they are likely to be more comfortable with its use. This acceptance is referred to as social licence. Social licence is dynamic and the level of acceptance can change over time, or indeed be suddenly lost. It is particularly dependent on the extent of trust the subjects hold in the data user, and their acceptance of the particular data uses. (Data Futures Partnership (July 2016) Exploring Social Licence)

Socio-technical: A socio-technical system or approach refers to the inter-relation of social and technical factors, systems and principles that lead to the production and use of a product. Sociotechnical elements could span physical infrastructure, like software and hardware, but also social and cultural factors and motivations. The example of a car is helpful: a car consists of an engine, computer system, steel frame, interior fittings, but once on the road, the person responsible for the car is required to observe social factors including road laws, road infrastructure and norms of driving.

Supervised learning: The process of training a model using training data that is labelled. For example, training a classifier to tell the difference between apples and oranges using training data made up of pictures labelled “an apple” or “an orange”.

Supervised machine learning: A form of machine learning that is trained on labelled data only.

Supplier: A person, business, company or organisation that supplies or can supply goods or services or works to an agency.

Test data: Data used to measure the performance of an algorithm, such as its accuracy or efficiency.

Training: The process used to create a model.

Training data: The set of data used in the training process to create, train or build an algorithm or machine learning model so it can accurately predict outcomes, find patterns or identify structures within the training data.

Transparency: The extent to which information regarding an algorithm or AI system is made available to Impacted People, including if one is used and the role it plays. It implies openness, comprehensibility and accountability in the way algorithms function and make decisions.

Te ao Māori: Te ao Māori acknowledges the interconnectedness and interrelationship of all living and non-living things via a spiritual, cognitive, and physical lenses. This holistic approach seeks to understand the whole environment, not just parts of it. There is no one Māori world view, in as much as there is no one New Zealander world view. The term is sometimes incorrectly interchanged with the term ‘mātauranga’. Mātauranga refers to soundly based knowledge and how it is attained. (This definition comes from Te Mātāpunenga)

Unfair outcomes: Used in the AIA to refer to potential bias, discrimination, and other unfair, unintended or unexpected outcomes.

Unsupervised learning: The process of training a model using training data that is unlabelled. For example, training an AI system to tell the difference between different kinds of vegetables using training data made up of unsorted and unlabelled pictures of vegetables.

Unsupervised machine learning: A form of machine learning that is trained on unlabelled data only.

User: A person who is intended to or will use the algorithm once deployed.

Variable: Any characteristic, number, or quantity that can be measured or counted. A variable may also be called a data item. Age, sex, business income and expenses, country of birth, capital expenditure, class grades, eye colour and vehicle type are examples of variables.

Appendix 1: Algorithm Charter

This Charter demonstrates a commitment to ensuring New Zealanders have confidence in how government agencies use algorithms. This Charter is one of many ways that government is demonstrating transparency and accountability in the use of data. However, it cannot fully address important considerations, such as Māori Data Sovereignty, as these are complex and require separate consideration.

Commitment

Our organisation understands that decisions made using algorithms impact people in New Zealand. We commit to making an assessment of the impact of decisions informed by our algorithms. We further commit to applying the Algorithm Charter commitments as guided by the identified risk rating.

Algorithm Charter commitments

TRANSPARENCY

Maintain transparency by clearly explaining how decisions are informed by algorithms. This may include:

- Plain English documentation of the algorithm
- Making information about the data and processes available (unless a lawful restriction prevents this)
- Publishing information about how data are collected, secured and stored.

PARTNERSHIP

Deliver clear public benefit through Treaty commitments by:

- Embedding a Te Ao Māori perspective in the development and use of algorithms consistent with the principles of the Treaty of Waitangi.

PEOPLE

Focus on people by:

- Identifying and actively engaging with people, communities and groups who have an interest in algorithms, and consulting with those impacted by their use.

DATA

Make sure data is fit for purpose by:

- Understanding its limitations
- Identifying and managing bias.

PRIVACY, ETHICS AND HUMAN RIGHTS

Ensure that privacy, ethics and human rights are safeguarded by:

- Regularly peer reviewing algorithms to assess for unintended consequences and act on this information.

HUMAN OVERSIGHT

Retain human oversight by:

- Nominating a point of contact for public inquiries about algorithms

- Providing a channel for challenging or appealing of decisions informed by algorithms
- Clearly explaining the role of humans in decisions informed by algorithms.

Appendix 2: DPUP principles

The Data Protection and Use Policy (DPUP) is based on the following five key principles.

- **He Tāngata** - Focus on improving people's lives — individuals, children and young people, whānau, iwi and communities. This incorporates privacy concepts such as data minimisation, purpose specification, and the creation of positive outcomes from data use.
- **Manaakitanga** - Respect and uphold the mana and dignity of the people, whānau, communities or groups who share their data and information. This incorporates recognition of diverse cultural perspectives about data, and requires meaningful partnership with affected service users.
- **Mana Whakahaere** - Empower people by giving them choice and enabling their access to, and use of, their data and information. This incorporates privacy concepts such as meaningful transparency, consent, and subject access and correction rights.
- **Kaitiakitanga** - Act as a steward in a way people understand and trust. This incorporates privacy concepts such as data protection (security), accountability, and privacy breach notification.
- **Mahitahitanga** - Work as equals to create and share valuable knowledge. This incorporates sharing data in ways that decrease the burden on service users and ensure the best outcomes for people and their communities, and also ensuring that de-identified data can be used for research and evaluation (though note specific open data risks discussed below).

The Data Protection and Use Policy (DPUP) was developed by Toi Hau Tāngata – Social Wellbeing Agency and implemented by Te Tari Taiwhenua – Department of Internal Affairs.

Appendix 3: References

Title	Organisation	Link
Data Protection and Use Policy		https://www.digital.govt.nz/standards-and-guidance/privacy-security-and-risk/privacy/data-protection-and-use-policy-dpup/
Trustworthy AI in Aotearoa: AI Principles 2020	AI Forum	https://aiforum.org.nz/wp-content/uploads/2020/03/Trustworthy-AI-in-Aotearoa-March-2020.pdf
Principles for the Safe and Effective Use of Data and Analytics	Privacy Commissioner and Stats NZ	https://www.stats.govt.nz/assets/Uploads/Data-leadership-fact-sheets/Principles-safe-and-effective-data-and-analytics-May-2018.pdf
Māori Data Sovereignty Charter and principles	Te Mana Raraunga	https://www.temanararaunga.maori.nz/tutohinga
Māori Data Governance Model	Te Kāhui Raraunga	https://www.kahuiraraunga.io/_files/ugd/b8e45c_a5b7af8b688c4cd9b7583775c27da52e.pdf
Ngā Tikanga Paihere: a framework guiding ethical and culturally appropriate data use	Stats NZ	https://data.govt.nz/assets/data-ethics/Nga-Tikanga/Nga-Tikanga-Paihere-Guidelines-December-2020.pdf
Government use of AI in New Zealand	University of Otago; Law Foundation of New Zealand (Colin Gavaghan, Alistair Knott, James Maclaurin, John Zerilli, Joy Liddicoat)	https://www.otago.ac.nz/__data/assets/pdf_file/0027/312588/https-wwwotagoacnz-caipp-otago711816pdf-711816.pdf

The use of algorithms in the New Zealand public sector	NZ Law Journal [2019] NZLJ 26	https://www.otago.ac.nz/__data/assets/pdf_file/0026/332981/liddicoat-j-gavaghan-c-knott-a-maclaurin-j-and-zerilli-j-the-use-of-algorithms-in-the-new-zealand-public-sector-a-preliminary-assessment-2019-new-zealand-law-journal-february-27-718754.pdf
The lessons learned from governance of Covid algorithms	Published in the Journal of the Royal Society of New Zealand	https://www.tandfonline.com/doi/full/10.1080/03036758.2022.2121290
Algorithm Information Request	New Zealand Algorithm Hub	https://assets.ctfassets.net/6uj5rbssnusx/1rx38jgFIHgef9O2jVdysL/6aebedee37267e95397c69865fff8d9c/Template_-_Algorithm_Information_Request.pdf
Algorithm Impact Assessment	Government of Canada	https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html
Algorithm Impact Assessment in Healthcare	Ada Lovelace Institute (UK) and NHS AI Lab	https://www.adalovelaceinstitute.org/project/algorithmic-impact-assessment-healthcare/
The proposed EU AI Act	European Commission	https://artificialintelligenceact.eu/
Ethics Guidelines for Trustworthy AI	High-Level Expert Group on Artificial Intelligence, European Commission	https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai
Algorithmic Transparency Recording Standard & Guidance	Central Digital and Data Office and Centre for Data Ethics and Innovation	https://www.gov.uk/government/publications/guidance-for-organisations-using-the-algorithmic-transparency-recording-standard
Guidance on AI and data protection	UK Information Commissioner's Office	https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection/
Data protection requirements when using biometric data	UK Information Commissioner's Office	https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/guidance-on-biometric-data/data-protection-requirements-when-using-biometric-data/

The OECD Artificial Intelligence (AI) Principles	OECD.AI Policy Observatory	https://oecd.ai/en/ai-principles
Catalogue of Tools & Metrics for Trustworthy AI	OECD.AI Policy Observatory	https://oecd.ai/en/catalogue/tools
AI Risk Management Framework	US National Institute of Standards and Technology (NIST)	https://www.nist.gov/itl/ai-risk-management-framework
Unpacking AI Procurement in a Box: Insights from Implementation	World Economic Forum (2022)	https://www.weforum.org/reports/ai-procurement-in-a-box/ https://www3.weforum.org/docs/WEF_AI_Procurement_in_a_Box_AI_Government_Procurement_Guidelines_2020.pdf https://www3.weforum.org/docs/WEF_AI_Procurement_in_a_Box_Project_Overview_2020.pdf https://www.weforum.org/whitepapers/unpacking-ai-procurement-in-a-box-insights-from-implementation/
Responsible AI Impact Assessment Template	Microsoft	https://blogs.microsoft.com/wp-content/uploads/prod/sites/5/2022/06/Microsoft-RAI-Impact-Assessment-Template.pdf
Responsible AI Guide	Microsoft	https://blogs.microsoft.com/wp-content/uploads/prod/sites/5/2022/06/Microsoft-RAI-Impact-Assessment-Guide.pdf
Automated Decision-Making in the Public Sector	Algorithm Watch	https://algorithmwatch.org/en/adms-impact-assessment-public-sector-algorithmwatch/
Algorithmic Accountability for the public sector	Ada Lovelace Institute; AI Now Institute; Open Government Partnership	https://www.opengovpartnership.org/wp-content/uploads/2021/08/executive-summary-algorithmic-accountability.pdf

ELI Model Rules on Impact Assessment of Algorithmic Decision-Making Systems Used by Public Administration	European Law Institute	https://www.europeanlawinstitute.eu/projects-publications/completed-projects/ai-and-public-administration/
AI: An accountability framework for Federal Agencies and Other Entities	US Government Accountability Office	https://www.gao.gov/products/gao-21-519sp
Algorithmic Accountability for the Public Sector	Open Government Partnership	https://www.opengovpartnership.org/documents/algorithmic-accountability-public-sector/
Checkmate Humanity – The how and why of Responsible AI	Sam Kirshner, Richard Vidgen and Catriona Wallace	https://checkmatehumanity.com/
Implementing Australia's AI Ethics Principles report	CSIRO, Gradient Institute	https://www.csiro.au/en/work-with-us/industries/technology/national-ai-centre/implementing-australias-ai-ethics-principles-report
Human Rights and Technology	Australian Human Rights Commission	https://humanrights.gov.au/our-work/technology-and-human-rights/publications/final-report-human-rights-and-technology

Te Kāwanatanga o Aotearoa
New Zealand Government