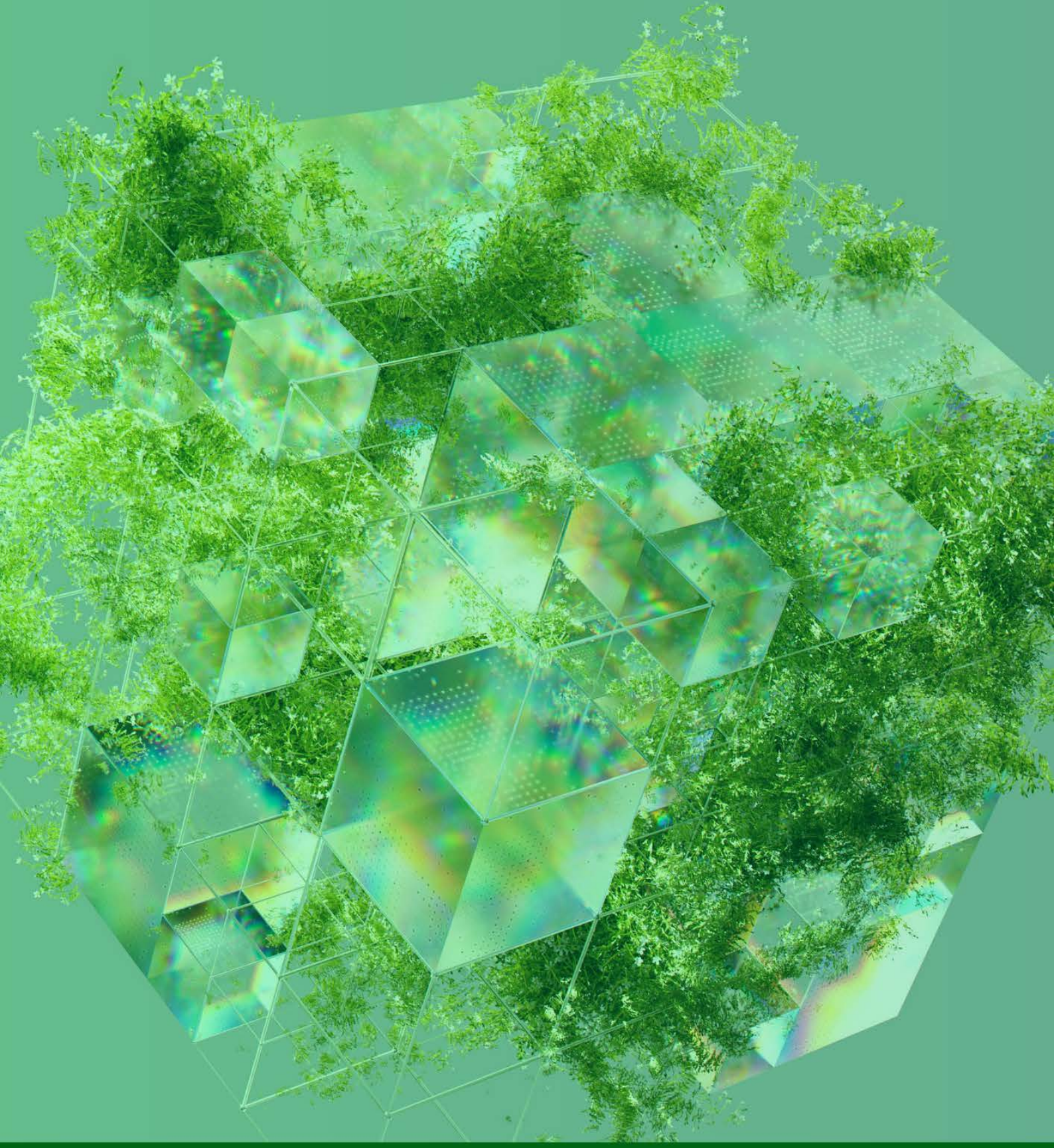


AI Forum
New Zealand
Te Kāhui Atamai Iahiko o Aotearoa

Explainable AI – building trust through understanding

A whitepaper from the AI Forum of New Zealand

November 2023

Acknowledgements

The AI Forum would like to acknowledge the XAI working group comprising subject matter experts from across Aotearoa New Zealand including: working group lead and Executive Council member Matt Lythe, Executive Council members Gabriella Mazorra de Cos and Maria Mingallon, AI Forum Executive Director Madeline Newman, AI Forum Chairperson Megan Tapsell, and working group members Dr Christopher Galloway, Dr Andrew Lensen, David Knox, Sarah Auvaa and Dr Kaushalya Kumarasinghe. We would also like to thank Sara Cole Stratton, Karaitiana Taiuru from, 'Te Kāhui Māori Atamai Iahiko', the Māori Advisory Panel in AI for guidance in embedding Te Ao Māori in this work and finally the Tech Alliance editorial team for assisting with the publication of this document.

About the AI Forum

The Artificial Intelligence Forum of New Zealand (AI Forum) is a purpose-driven, not-for-profit, non-governmental organisation (NGO) that is funded by members. We bring together New Zealand's community of artificial intelligence technology innovators, end users, investor groups, regulators, researchers, educators, entrepreneurs and interested public to work together to find ways to use AI to help enable a prosperous, inclusive and thriving future for our nation. The Forum holds an evidence based approach, focusing on addressing challenges to realise opportunities.

Table of Contents

Executive Summary

Artificial intelligence (AI) has shown great potential in many real-world applications, for example, clinical diagnosis, self-driving vehicles, robotics and movie recommendations. However, it can be difficult to establish trust in these systems if little is known about how the models make predictions. Although methods exist to provide explanations about some black box models these are not always reliable and may be even misleading.

Explainable AI (XAI) provides a meaningful solution to this dilemma in instances where it may be important to explain why an AI model has taken certain actions or made recommendations. These models are inherently interpretable, offering explanations that align with their computations, resulting in improved accountability, fairness and less bias. However, explainable models can also be less capable or versatile and may decrease model accuracy when compared to more complex, less transparent models.

The demand for explainability varies with the context. The more critical the use case, the greater the need for interpretability. For example, the need for interpretability in an AI based medical diagnosis system would be significantly higher compared to one used for targeted advertisements. In Aotearoa New Zealand there are already excellent examples of XAI including in health, justice and the environment. The potential for many more systems is substantial, especially when AI decisions affect people or communities in a significant way.

Matt Lythe
Chair of the Working Group on Explainable AI
AI Forum New Zealand Executive Council Member

Introduction

Artificial Intelligence (AI) is impacting all parts of the Aotearoa New Zealand economy, creating transformational opportunities for business, Government and society. However, most advanced AI models, including deep learning neural networks, often operate as 'black boxes', with internal workings that are invisible to the user. The user can provide an input, receive an output, but cannot examine the system's code or logic that provided it. This makes it challenging for people to understand the system's decision-making processes. As adoption of AI becomes mainstream and penetrates deeper into our everyday life, people are becoming concerned regarding the use of their data and the wider risks of AI.

In Aotearoa the incorporation of Tikanga Māori[1] into AI deployment is paramount. AI systems must respect principles embracing a Te Tiriti lens for any algorithmic development. Deployments that use or produce Māori data, or decision-making about Māori should be adopted in conjunction with Māori Data sovereignty principles. Examples of recent ethical principles developed for AI use include[2]:

- *Principle 1: Tino Rangatiratanga*
  All AI systems will embed Māori leadership, decision-making and governance at all levels of the systems life cycle from inception, design, release to monitoring. Māori should be engaged at an early stage to co-develop uses of AI and ensure that Māori data is stored appropriately.

- *Principle 2: Equity*
  AI systems will achieve equity outcomes for Māori (individuals and collectively) across the life course and contribute to Māori development. This involves businesses and employees who are accountable to Māori in how AI models are used with Māori data and outputs that impact Māori individually and collectively, and in the active building of capacity of the Māori AI and tech workforce.

- *Principle 3: Active Protection*
  Requires free, prior, and informed consent (FPIC) for the use of Māori data in AI development, with robust procedures in place to prevent biases or predictions that stigmatise or harm Māori.

[1] Tikanga Māori refers to the concept of incorporating practices and values from mātauranga Māori (Māori knowledge or wisdom).

[2] Source: Kariatiana Taiuru; http://www.taiuru.maori.nz/AI-Principles

- Principle 4: Mana Whakahaere

  Effective and appropriate stewardship or kaitiakitanga over AI systems is required. It is recognised that Māori data is a taonga and subject to Māori Data Sovereignty principles determined by Te Tiriti. A deep understanding of the source and intended use of data is required, so that it is not repurposed without permission or in a way that will diminish the mana of Māori.

- *Principle 5: Mana Motuhake*
  Requires that tikanga (practices) are followed throughout AI development and deployment, with Māori deciding what data and data uses are controlled (tapu) or allowed (noa).

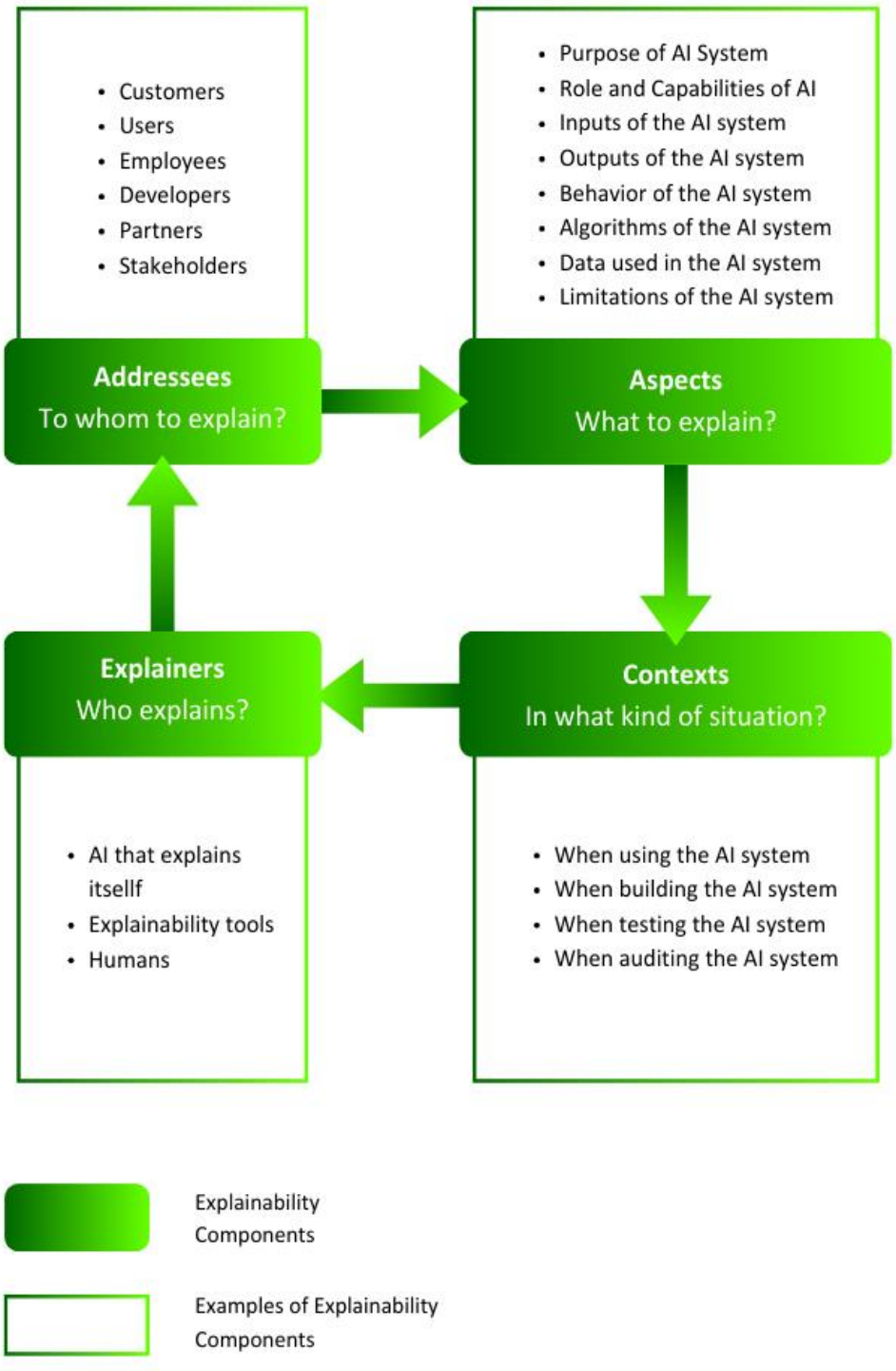
- *Principle 6: Tapu/Noa; Cultural safe practices*
  No AI will be culturally unsafe or break the rules of Tapu and Noa.

When deploying AI systems, achieving these principles will often require an AI model that is *explainable.* Throughout this whitepaper, we reference these principles to provide examples of how they are applicable to explainable AI. Nevertheless, this is not an extensive discussion of how to meet *Te Tiriti* obligations using XAI – companies should develop their own internal polices and meaningfully engage Māori stakeholders at an early stage.

But what does 'explainable' mean?

Explainability has multiple inter-related components, including:

- the explainers (is it a human or the AI?),
- addressees (who is it being explained to – developers, users, or customers?), contexts (at what stage of development/deployment does the explanation take place?)
- and aspects (what is being explained?).



Within AI systems, explainability aims to answer questions about the decision-making, enabling users to understand the rationale behind their outputs. Explanations help end-users gain an understanding of how AI systems work and address questions and concerns about their behaviour. Recently, AI researchers have recognised explainable AI (XAI) is

necessary for trustworthy AI, where the level of required explainability depends on the use.

If their AI system (purchased or developed) does not perform correctly, governments and companies are at risk of unintended consequences and reputational damage. XAI provides confidence that the system can accurately make the right decisions. This increases user trust and societal confidence that the system is operating ethically, without manipulation and bias.

In resonance with this, the United Kingdom (UK) Information Commissioner's Office and the Alan Turing Institute have published guidance, *Explaining decisions made with AI*, outlining six primary types of AI decision explanations:

I. Rationale – reasons that led to the decision;
II. Responsibility – who is involved in the AI system's development and management, and who to contact to request human review of a decision;
Data – what data has been used for a decision, and how;
III. Fairness – what steps have been taken in the AI design and implementation to ensure unbiased and fair decisions;
IV. Safety and performance – how the AI system's accuracy, reliability, security, and robustness is maximised; and
V. Impact – how the impacts of the decision on individuals and society have been considered and are monitored.

In considering which explanation(s) is appropriate for a particular AI model, it recommends evaluation of:

- the context and area in which the AI model is deployed,
- the degree and nature of the impact of the AI decisions on individuals,
- the type and sensitivity of the data used or created,
- how urgent it is for the individual to understand/ make choices based on the AI model outcome, and
- the audience the explanation is intended for, including their degree of expertise in the subject area and understanding of AI.

When is explainability important?

When the decision has an impact on people and communities.

This whitepaper investigates explainability in AI, with a particular focus on the field of Machine Learning (ML) which has produced the majority of recent advances and renewed interest in AI. We look at why and when explanations are useful and when they may not be needed. We will discuss the relationship between AI performance and explainability and the benefits of interpretability including better error detection, enhanced user understanding, reducing bias and increasing trust.

We also present several XAI projects from Aotearoa New Zealand which provide context and guidance on this rapidly developing field for developers and users.

Understanding Models

In this section, we introduce some common terms used in the field of explainable AI. There are many sources of definitions including academic papers in different fields, blog posts from practitioners, documentation from popular code libraries, and documentation from industry. Many of these terms are defined differently in different sources. Further, some sources provide precise definitions of these terms while others use some of these terms interchangeably.

In the realm of explainable AI, two core concepts often surface: **Explainability** and **Interpretability**. Both terms are vital to our understanding, but they are distinctly nuanced. Before diving deep, it is imperative to establish a foundational understanding of some prevalent terms and the role of 'models' in the AI context.

Within AI, a model is a mathematical representation or computational system formulated through specific algorithms to make data-driven predictions or decisions. When discussing machine learning, a model is a predictive algorithm shaped by training on input data. Artificial intelligence applications, for example ChatGPT, are nearly always built using these models.

Explainability

Explainability of AI refers to the ability to describe a model's decision-making process in a way that is relatable to humans. Explainability is profoundly contextual. For instance, an AI specialist's level of understanding is vastly different from an ecologist employing AI for research or a consumer interacting with an AI generated advertisement.

It is crucial that, when choosing an XAI model, we consider the recipient of the explanation – the target audience.

This contextualisation is discussed throughout this white paper — in certain situations, there may be no need at all 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). Even within a specific use case, individual requirements will also vary. For example, one patient may prefer the most accurate and unexplainable diagnostic AI system, whereas another may refuse to accept an AI decision that they cannot follow in detail.

This contextualisation is crucial in honouring principles of Māori data sovereignty – for Māori to have *Tino Rangatiratanga* over the use of Māori data, the functionality of an AI model must be explained in a way that can be understood according to Māori worldviews. Explainable AI is one piece of the puzzle in demonstrating the *Active Protection* to those whose data is being utilised.

Interpretability

Interpretability delves deeper into grasping the intricate mechanics of how a model makes its decisions. A completely interpretable model allows humans to replicate its decisions using tools like spreadsheets or even manual calculations.

There are two key types of interpretability:

- **Local Interpretability** refers to the ability to reproduce a model's decisions for a specific input. For example, a single prediction.
- **Global Interpretability** encompasses a broader scope, ensuring the model's entire decision-making mechanism can be translated into human-readable formats, be it decision trees, mathematical equations, or code segments.

The table below provides further comparisons between these two terms.

Aspect	Explainability	Interpretability
Definition	Articulating the model's decisions in human-relatable terms	Deciphering the model's decision-making mechanics
Focus	***Why*** a decision was made	***How*** a decision was made
Audience	General public, end users, stakeholders	Data scientists, AI researchers, technical experts
Importance in Scenarios	High-stakes decisions, trust-building	Model refining, debugging, ensuring fairness
Example	Explaining why a loan was denied	Understanding the nodes in a decision tree for a loan application

Transparency, glass box and black box models are additional terms that will be used in our discussion.

Transparency

Transparency in AI refers to the ability to understand and explain how an AI system makes decisions or predictions. It goes beyond the concepts of interpretability and explainability; providing clear insights into the underlying algorithms, data used, and the reasoning behind the AI's outputs, ensuring accountability, fairness, and trustworthiness in AI applications. Transparency enables individuals to make informed decisions regarding their data and privacy *(Tino Rangatiratanga*). When done well, transparency also provides people with increased confidence in the organisation. It also helps inform constructive public debate about the benefits and risks of data use and AI.

The content and level of detail concerning transparency are contextual, hinging on various factors such as the model's structure, the scenario it's deployed in, the type and sensitivity of any personal information, and the significance and impact of the model's outcomes on individuals.

To ensure the appropriate level and type of transparency, thorough planning is required prior to deploying any AI model. This includes:

- AI model selection and development**:** the chosen model and its development pathway should facilitate the required levels and types of transparency, especially in cases where public trust is pivotal.
- Data inputs and outputs: the types, sources, and collection methodologies of any personal information used in the model need to be clearly documented *(Mana Whakahaere*). Additionally, any new personal information data points generated by the model, along with how that information will be used and shared, must be documented. Undertaking a Privacy Impact Assessment can aid in this process.
- Implementing and maintaining controls: it is crucial to ensure that controls (for example, checking data sets and outputs for potential bias) referenced in the transparency description are robustly implemented and maintained *(Active Protection*).
- Model governance: processes must be established for determining which model alterations necessitate an update to the transparency description. This requires assessing the impact (and risks) of any proposed changes, documenting approved changesand updating the transparency description accordingly. This includes deciding

whether individuals should be proactively informed of the changes or if updating the description suffices *(Equity*).

- Monitoring for feedback: channels should be created for receiving and addressing feedback about the transparency description (.for example, do people understand it?) and about the model and data (for example, are there concerns about whether the data is representative of individuals impacted?).
- Accountability: responsibility for all these requirements must be designated in writing to relevant roles, with ensuring that individuals possess the necessary information, training, capability, and resources to deliver them *(Equity*).

We recommend that transparency needs should be identified and clearly stated alongside the kaupapa when starting the development of any AI model. This includes determining the assurances and controls necessary to gain the confidence of stakeholders, including those who may be impacted by the model. This allows organisations to incorporate the necessary levels of explainability and supporting processes into their AI models and establish public trust through clear transparency descriptions.

Additionally, relevant areas of the organisation should be engaged to provide support – for example engaging teams with communications expertise, to help draft the transparency description, and ensuring legal review to support legislative compliance. Ensuring these teams have sufficient training on AI is also crucial.

Ensuring that your transparency description is readily available to the intended audience is critical. For example, is it accessible to people who may have a disability or people who use assistive technologies? The Web Accessibility Guidance project provides guidance on accessibility. Similarly, consider the languages that your audience will prefer to use to read the transparency description, and how to meet their needs.

Glass box models

Also termed as 'white box' in some contexts, glass box AI models provide interpretability and transparency. In glass box models, for example, linear regression, logistic regression, decision trees or LIME (Local Interpretable Model-Agnostic Explanations), the underlying mechanisms for generating predictions can be understood allowing users to understand the decision-making process and gain insights into the underlying rules and variables.

Black box models

Conversely, by their inherent design, black box models are more intricate and less discernible than their glass-box counterparts. This may stem from their complexity, for example, deep neural networks or generative AI like large language models (LLMs or owing to proprietary or closed-source constraints. The intricacy of these models, often with millions of adjustable parameters, impedes human interpretation. Other models including random forest (RF) and gradient boosted machines (GBM), consisting of numerous decision trees, are considered black box models due to their scale and complexity.
However, the label doesn't render these models completely unexplainable. Various tools and techniques have been developed to glean insights into their decisions, which are discussed later. Such models can still maintain transparency by making known their training datasets, architecture specifics and modelling goals. This highlights their origin and purpose, fostering greater trust and understanding.

Explainable Models

It is widely held that black-box models are necessary for the highest predictive performance. However, this is often not true, particularly if the data are structured, with a good representation in terms of naturally meaningful features.

In this section, we summarise the most popular explainable approaches and explore their potential in replacing more opaque AI models.

1. **Linear regression** is a statistical method that models the relationship between a dependent (output) and one or more independent (input) variables by fitting a linear equation.
    a. **Explainability**: Linear regression quantifies the relationship between variables. Each variable's coefficient signifies its impact on the outcome, making it easier to understand and convey.
    b. **Replaces**: Black box models (like deep neural networks), especially in tasks where variables have linear relationships. For example, when predicting house prices based on features including area, number of rooms, and locality, a simple linear equation may suffice, rather than a complex neural network.
2. **Generalised additive models (GAMs)** are an extension of linear regression. Instead of strictly linear relationships, GAMs allow for non-linear relationships (splines) as well. They are *additive* models, as they add together multiple individual relationships to get a final prediction.
    a. Explainability: Each relationship found by a GAM can be interpreted individually. This, combined with the additive approach, provides a model that can both perform highly and be explained.
    b. Replaces: Non-linear black box models like support vector machines (SVMs) or neural networks in settings where capturing and interpreting complex relationships is necessary. They're particularly useful in contexts such as econometric studies where understanding factor influences is just as crucial as prediction accuracy.

3. **Decision trees** build a tree that makes a decision at each step based on the input until reaching a final answer.
    a. **Explainability:** The structure visually shows how decisions are made, making it easy for stakeholders to understand the decision path. It provides an explicit path of logic, illustrating exactly how a decision was reached.
    b. **Replaces:** Complex ensemble models like random forests or gradient boosted trees in scenarios where a single, interpretable decision path is desired over a (higher-performing) consensus of many trees. They are especially useful in areas like clinical decision support where each diagnosis requires a clear rationale.

4. **Genetic programming** automatically constructs computer programs to perform a task. The resulting models, often represented as tree structures, share similarities with decision trees. The evolutionary-based training process (using concepts like selection, crossover, and mutation) provides transparency into the model's creation.
    a. **Explainability:** While more complex than standard decision trees, they are still much more interpretable than black box models.
    b. **Replaces:** Black box optimisation or learning algorithms, especially in tasks where the objective is non-differentiable. For example, in supply chain optimisation, stakeholders may prefer a solution that, while potentially less efficient than one derived from a black box model, provides clear decision logic, aiding broader strategic plans.

5. **Rule-based systems** (for example, learning classifier systems) learn a series of 'if-then' rules. Each decision is the direct result of evaluating these rules. These are often used in expert systems and business settings where rationale behind each decision is essential.
    a. **Explainability:** Stakeholders can trace every decision back to a specific rule, offering complete transparency.
    b. **Replaces:** Proprietary or closed-source models where the internal decision-making process isn't revealed. For instance, in regulatory banking sectors, every loan approval or denial can be mapped back to a specific rule, facilitating audits and compliance checks.

6. **Bayesian networks** are graphical models that represent the probabilistic relationships among a set of variables. Bayesian networks can capture intricate relationships and provide a holistic view of variable interdependencies.
    a. **Explainability:** The graphical nature depicts variable interactions, and given evidence, the network provides reasoning via inferred probabilities. This allows stakeholders to see both the outcome and the probabilistic logic behind it for different inputs, including 'what-if' scenarios.
    b. **Replaces:** Black box probabilistic models in fields like medical diagnosis, where understanding the chain of reasoning and causal relationships is crucial.

When considering problems that have structured data with meaningful features, there is often no significant difference in performance between more complex classifiers and much simpler classifiers.

Bias and Trust

Explainability is a pivotal tool not only for interpreting AI decisions, but also for identifying biases and fostering trust. Explainability provides insights into the decision-making process, but to ensure trustworthy operations, the inherent challenges of bias must be explored. This section examines the multifaceted nature of trust and its nuances.

Bias

Implementing AI in real-world applications presents numerous challenges, with bias in machine learning being paramount. This bias can influence results, leading to varied outcomes for different demographics. Identifying, assessing and mitigating these biases is crucial to achieve trust in AI systems. Identifying and removing bias is also a key factor in demonstrating *Active Protection* and ensuring Māori collectively benefit from the use of AI in Aotearoa (*Equity*).

In 2023, Bloomberg Technology highlighted the biases in Generative AI, showing that the stable diffusion model exaggerated racial and gender disparities. Analysis revealed that the AI consistently associated lighter skin tones with high-paying jobs and darker skin tones with roles such as fast-food worker and social worker.

"Every part of the process in which a human can be biased, AI can also be biased", said the AI Center for Policing Equity. This indicates the clear links between societal biases and embodied AI biases.

Types of bias:

- **Human cognitive bias**: Inherent biases in human thinking and decision-making processes. These include:
    - **Automation bias**: Preferring automated system suggestions, even when contradictory information is accurate.
    - **Confirmation bias**: Favouring AI predictions that align with pre-existing beliefs.
- **Societal and systemic bias**: these reflect societal viewpoints and can be deeply rooted in systems, cultures, or organizations. They emerge when AI models learn from or magnify existing societal biases.
- **Computational bias**: Errors stemming from the foundational assumptions of a model or the underlying data.

- **Data Bias**: Biases originating from the data itself, which can arise from:
    - **Statistical bias**: Including selection, sampling, and non-response bias.
    - **Labelling issues**: Errors in data labelling or in the labelling process itself.
    - **Sampling concerns**: Using non-representative samples.
    - **Incomplete data**: Missing features or labels.

Trust

As we've established, biases in AI can affect its decision-making processes. However, it's equally important to understand that the presence or perception of such biases can deeply affect trust in AI systems. Trust in AI is more than just being told the model is accurate; it is a willingness to accept vulnerability, relying on system outputs and sharing data, grounded in the positive expectation of the system's operation. Trust spans multiple facets including reliability, robustness, fairness, and ethics. It is influenced both by the system's perceived performance and its explainability. Moreover, the degree of transparency in AI's development and the active participation of diverse stakeholders play a significant role.

A 2023 study from the University of Queensland illuminated the complexity of public sentiment toward AI. Notably, while 61 percent of respondents voiced apprehension about trusting AI, 82 percent were aware of the technology. This dichotomy reveals a gap in understanding, with half of those familiar with AI unsure about its operations, highlighting explainability's importance. Another revealing statistic is only 50 percent believed that the benefits of AI surpassed its risks, 75 percent indicated they would place more trust in AI if ethical guidelines were established; most respondents believed there are insufficient safeguards in place to govern AI. Further bolstering this, 80 percent saw the merit in system accuracy and reliability monitoring, and 68 percent believed that adhering to AI ethics certifications would be beneficial.

Regulatory Considerations

Although New Zealand lacks specialised AI legislation of the European Union, there are still several regulatory frameworks that must be followed when developing and deploying AI.

The **New Zealand Privacy Act 2020** stipulates that AI models handling personal information must meet specific standards for collection, use, and disclosure of personal information. There is also a requirement for transparency; organizations must inform individuals about the purpose of data collection and the data's usage (Privacy Act Principle 3). Organisations may only use collected data for directly related purposes, unless they have the individual's explicit authorisation or are covered by a Privacy Act exception (Principle 11).

AI model explainability will often be critical to enable Privacy Act compliance, particularly when the model's outcomes impact individuals. Given the pace of AI development, organisations must stay up to date on regulations. The Office of the New Zealand Privacy Commissioner provides regularly updated resources setting out their expectations, including recent Generative AI guidance:

> ***Be transparent***
> *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.*

Another cornerstone is the **Human Rights Act 1993**, which prohibits discrimination on grounds such as gender and race. To ensure compliance with this Act, organisations using AI for processes such as recruitment (for example to create a shortlist of applicants) will greatly benefit from an XAI approach that can show the AI model is fair and non-discriminatory. Providing this transparency to the job applicants would also help to allay concerns of discrimination.

On the global stage, regulations are evolving rapidly, providing context for how New Zealand legislation may itself develop. The **European Union's General Data Protection Regulations (GDPR)** mandates that

individuals be informed about automated decisions, especially those without human intervention. They must also receive insight into the AI's logic and potential consequences (as per Articles 13 and 14). Furthermore, the GDPR mandates the use of Data Protection Impact Assessments for any data processing that could put individuals at high risk (Article 35) - these assessments will be much easier with an explainable model.

The forthcoming **EU Artificial Intelligence Act** includes more detailed provisions. It necessitates transparency for Generative AI, requiring acknowledgment when such AI is used to produce content. Additionally, it demands summaries of copyrighted data utilized for training. Systems categorised as high risk and limited risk, encompassing technologies like chatbots and emotion recognition systems, will be subject to mandatory transparency mandates that reflect their risk factors.

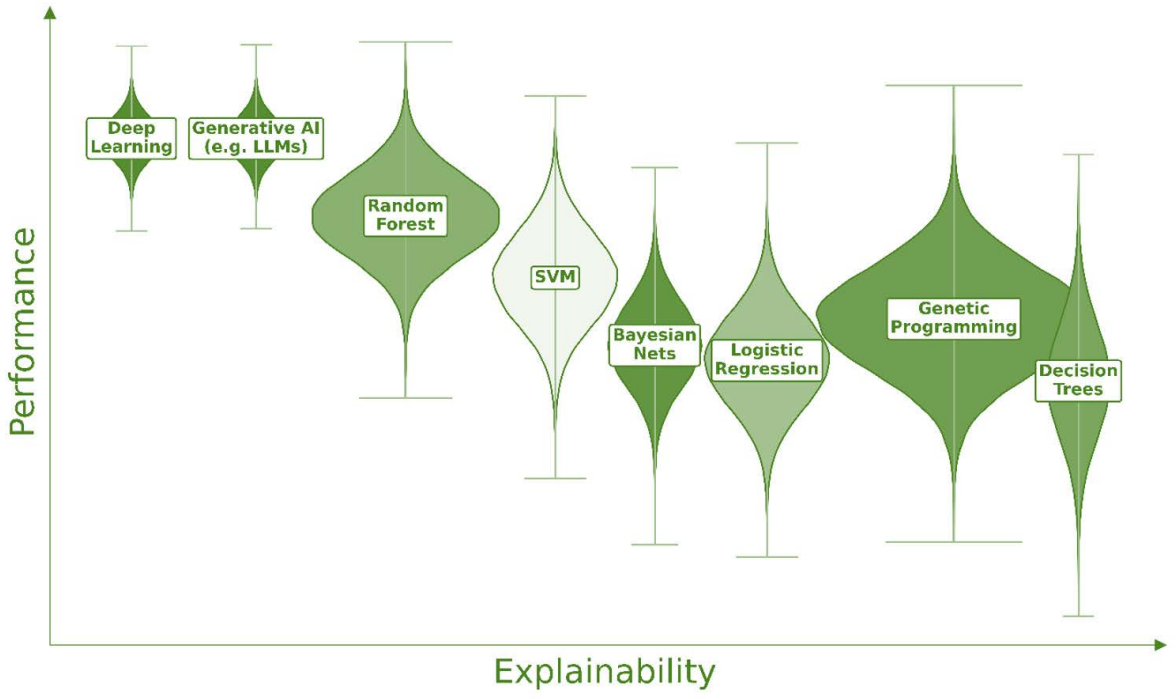
This regulatory wave isn't confined to Europe. Similar frameworks are emerging in Australia and various US states. New Zealand businesses operating internationally must be vigilant, as they may already be subject to some of these regulations. Moreover, increasing demands from clients, benchmarking organisations, and general global trends emphasise the escalating need for transparency in AI practices.

Balancing Explainability and Performance

Performance, a crucial aspect of AI, often benefits from complex models like deep neural networks. These models are adept at capturing intricate patterns in data, which in turn drives their accuracy. However, their complexity inherently makes them opaque, creating a black box scenario where the decision-making process is obscured.

Explainability seeks to demystify this process, ensuring AI decisions are understandable. Yet, simplifying models to enhance their explainability can sometimes diminish their capacity to capture data nuances, leading to reduced performance. This presents a fundamental trade-off: as one tries to maximize either performance or explainability, the other generally will suffer.

We illustrate this trade-off in the figure below for a selection of common and/or explainable machine learning models. For each model, we show its typical range of performance and explainability as a violin plot. Deep learning and Generative AI, for example, have consistently high performance but are also consistently unexplainable. In contrast, a simple model like a decision tree is generally very explainable, but its performance is heavily dependent on the data and application.



How well can we explain our most complex models using explainable techniques?

Complex models, such as deep neural networks (DNNs), are difficult to explain due to their enormous number of parameters. OpenAI's GPT-3, which gave us ChatGPT, has 175 billion parameters, and its successor GPT-4, is rumoured to be ten times more complex. However, it is possible to gain some sort of partial explainability. While DNNs may never be as interpretable as a simple decision tree, there may be situations where their superior performance means that a partially explainable model is acceptable. We highlight some approaches to explaining complex DNNs below.

Deep learning

Feature visualisation techniques create visual representations of the inner workings of a DNN. Each neuron in a neural network has an activation function, which produces a higher output when that neuron activates. By visualising the activations of different layers of a network, we can gain some insight into what the model is looking at. These techniques become less effective on very deep networks, where later layers are far from the input layer.

Saliency maps are another visualisation that looks at what parts of the *input* (for example an image for image classification tasks) were most influential in the model's decision-making. They can be quite sensitive to noise, which means they do not always represent causal relationships.

Local interpretable model-agnostic explanations (LIME) is a generic explainability method that can be applied to various ML algorithms, including DNNs. As a *local* method, it is useful for explaining why a specific model output (given a single model input). For example, in predicting house prices, LIME may tell us why a DNN valued your house at $750,000, but it will not help you understand the function of the model across all different inputs. LIME achieves this by building a simpler, interpretable model (for example a linear model) that approximately explains the output for a given input. While LIME can be powerful, it is computationally expensive and may not give accurate explanations for extremely complex models.

The above approaches can be used for all kinds of DNNs. There are also more tailored approaches to explaining specific types of DNNs, including generative and large language models.

Generative models

Generative AI models, which generate text, images, or other media based on a text prompt, have unique characteristics that allow them to be partially explained in more specific ways. These models produce a *latent space*, an internal numerical representation that represents both the text prompt and the output media. **Latent space analysis** looks at patterns in this space to try and understand what the model has learned. For example, small changes to the latent space should make small changes to the output media, providing insight into what this space represents. More advanced analysis can interpolate between points in this space: interpolating halfway between 'cat' and 'dog' would give an idea of what the model thinks a cat-dog may look like.

Model inversion techniques flip the model on its head: given an output media, what text prompt would have produced this output? These techniques provide insights into what features the model considers important (and any biases it may have) but are computationally expensive and not feasible in many cases. These approaches are also useful for uncovering potential privacy issues: there have been cases where model inversion has uncovered information such as phone numbers from the training data set.

We can also use counterfactuals to understand generative AI models. **Counterfactual analysis** uses 'What if?' questions to understand a model's function. For example, for the prompt 'a smiling cat', we may ask, 'What if we change *smiling* to *angry?'*. If the model makes a large change to the output (for example producing an image of an angry cat instead), then we can conclude that the model considers smiling an important factor in its decision-making.

Large language models

While LLMs are most commonly used for text generation (for example ChatGPT), they are also often used as *foundational models* for other downstream tasks such as text classification or sentiment analysis. As they are specifically designed for text inputs, there are explainability techniques specifically designed for LLMs.

Attention maps can be used in transformer-based models (GPT3/4) to visualise what part of the text input the model focuses on when producing an output. However, attention is not the same as explanation: a model may pay attention (in the technical sense) to a specific word even if that word is not crucial for the model's decision-making. The correlation between attention and importance in a model's decision-making process is not as strong as previously thought.

Recent research has begun to explore using **vector databases** or **retrieval augmentation** to enhance the explainability (and performance) of LLMs. These approaches combine more traditional document retrieval approaches with the power of LLMs. For example, if a doctor had a database of medical texts and diagnosis guidelines, the vector database could find the most relevant tests based on the input query. These would then be used to augment that query to the LLM, providing it with additional information for a more accurate answer. This approach is also more explainable: the doctor can understand what sources the AI system used, increasing their confidence in its findings.

XAI's unique challenges

There are several unique challenges that need to be considered. For example, the computational cost of some explainability techniques can be a significant hurdle. While insightful, methods like LIME or model inversion can be computationally intensive, making them less feasible for large-scale applications or extremely complex models.

Another challenge lies in the validation of explanations. Verifying whether an explanation accurately represents the model's decision-making process can be difficult, especially for complex models where the relationship between inputs and outputs can involve intricate interactions between numerous variables.

Lastly, the dependency on external resources, such as document databases in retrieval-augmented models, introduces another layer of complexity. The quality, relevance, and timeliness can significantly impact the model's performance and explainability.

Which industries could benefit from XAI?

Interpretable or explainable AI models offer a solution to a diverse array of problems across the economy and have the potential to have a dramatic impact across various industries. These models can be employed to address a range of challenges by providing transparent and understandable insights, while mitigating biases and enhancing trust in AI-driven systems. In this section we explore industries that will benefit and problems that can be successfully tackled using interpretable AI models.

Healthcare: Interpretable AI is poised to revolutionise healthcare by transforming the way medical professionals diagnose, treat and manage various health conditions. In diagnostics, interpretable AI models can assist doctors in analysing medical images, pathology data, and patient records, enhancing accuracy and efficiency. By explaining the reasoning behind their decisions, these models build trust and enable personalised treatment plans, leading to better patient outcomes. Interpretable AI also aids in drug discovery and development by identifying potential drug targets and predicting drug interactions. Additionally, these models can optimise resource allocation, streamline operations, and reduce costs.

Finance: Explainable AI can improve risk assessment, fraud detection and investment strategies. Fraud is a significant concern in the finance industry, and explainable AI can bolster efforts to detect and prevent fraudulent activities. By providing understandable reasons behind the identification of suspicious transactions or patterns, interpretable AI can aid investigators and analysts in identifying new fraud trends and adapting their strategies accordingly. Equally, investors often rely on AI-driven models to make investment decisions. Interpretable AI can provide clear explanations for investment recommendations, helping investors understand the underlying factors driving the predictions. This empowers them to make more informed choices and manage their portfolios effectively. In the insurance industry, explainable AI can help with risk assessment, premium calculations, and claims processing, fostering trust and accuracy.

Manufacturing: Interpretable AI will greatly benefit the manufacturing industry by enhancing quality control, optimising production processes, and reducing defects. Transparent AI models offer clear insights into factors affecting product quality, allowing manufacturers to make data-driven decisions. This fosters consistency in product output and efficient resource allocation. By understanding the reasoning behind AI-driven decisions, manufacturers can improve operational efficiency and overall product quality, leading to increased customer satisfaction and reduced production costs.

Cybersecurity: Interpretable AI has the potential to significantly bolster the cybersecurity industry by enhancing threat detection, response, and overall resilience against cyber-attacks. Transparent AI models can provide clear explanations for how they identify and analyse potential threats, enabling cybersecurity experts to understand the reasoning behind the AI's decisions and validate the accuracy of their predictions. This understanding allows security professionals to fine-tune the AI models, adapt strategies, and stay ahead of emerging threats. Interpretable AI also aids in root cause analysis, helping to identify vulnerabilities and weak points in systems, leading to targeted security improvements.

Transportation: Interpretable AI holds great promise for the transport industry, offering a range of benefits that improve safety, efficiency, and customer experience. In autonomous vehicles, interpretable AI can provide clear explanations for the decisions made during driving, enhancing trust and acceptance of self-driving technology. For traffic management and logistics, interpretable AI models can optimise route planning, resource allocation, and fleet management, leading to reduced congestion and lower operating costs. Additionally, in public transportation, interpretable AI can help optimise schedules and service routes, improving the overall commuting experience for passengers.

Government and public policy: Interpretable AI holds great promise in government and public policy by offering transparent and understandable insights into complex decision-making processes. For example, AI can assist policymakers in making informed decisions in a range of ways, including criminal justice, social welfare, education, and healthcare allocation. By providing clear explanations for its

recommendations, AI can help avoid biases and ensure fairness in policy decisions. It enables policymakers to understand the factors influencing outcomes, leading to more effective and equitable policies. Additionally, interpretable AI fosters accountability and allows for scrutiny by the public promoting greater trust in government initiatives. Ultimately,

Legal and compliance: Interpretable AI holds significant potential in the legal and compliance industry by providing transparent and understandable insights into complex legal processes. In contract analysis, legal research, and compliance monitoring, interpretable AI can assist legal professionals in making more informed decisions. By explaining the rationale behind its conclusions, AI models can assist lawyers and compliance officers in accurately validating and interpreting legal documents, saving both time and resources. Furthermore, interpretable AI aids in identifying potential biases in legal decisions, ensuring fairness and adherence to ethical standards.

Customer service and support: Interpretable AI offers significant benefits to the customer service industry by providing transparent and understandable insights into customer interactions. Using chatbots and virtual assistants, interpretable AI can deliver clear and accurate responses, increasing customer satisfaction and trust. Customers can better comprehend the AI's reasoning, leading to improved communication and personalised service. Interpretable AI also aids in analysing customer feedback and sentiment, enabling businesses to identify areas for improvement and enhance their products and services accordingly.

Human resources (HR): Interpretable AI offers valuable support to the HR industry by providing transparency and fairness in various processes. In candidate screening, interpretable AI models can help identify the reasons behind selections and rejections, ensuring unbiased hiring decisions. In employee performance evaluations, interpretable AI enables clear explanations for assessments, leading to improved feedback and development opportunities. Moreover, interpretable AI aids in workforce planning, providing insights into factors affecting employee turnover and productivity. This technology promotes diversity and inclusion by mitigating biases and enhancing understanding of HR decisions.

Environment. Interpretable AI holds significant potential to benefit the environment by enabling data-driven decision-making and fostering transparency in various environmental applications. In climate

modelling and prediction, interpretable AI models can provide clear insights into the factors influencing climate change, aiding scientists and policymakers in understanding and mitigating its impact. For environmental monitoring, these models can analyse data from sensors and satellites, identifying patterns of pollution, deforestation, and habitat loss, facilitating proactive conservation efforts. Interpretable AI also supports sustainable resource management by optimising energy consumption, water usage and waste reduction. By explaining the rationale behind its recommendations, interpretable AI promotes public awareness and trust in environmental initiatives, encouraging individuals and businesses to adopt eco-friendly practices. Ultimately, interpretable AI has the potential to play a pivotal role in addressing environmental challenges and creating a more sustainable future.

Interpretable models and transparency in Aotearoa New Zealand

In New Zealand there are several examples of explainable approaches already in use. These span both the public and private sectors and encompass a range of algorithmic approaches from statistically based models to AI. In this section we discuss several case studies.

In the public domain **Accident Compensation Corporation (ACC)** are using statistical models to approve some claims and populate information about claims. For example, the ACC Accident Description Service (ADS) searches the free text in the ACC45 claim form, looking for keywords that could help categorise the type of accident being claimed for, for example, rugby accident or fall. It uses statistical models to auto-populate certain data fields that are used for injury prevention, monitoring and reporting purposes as well as by the Actuarial team. The ADS information is not used as part of claims approval. Where there is not a statistically likely result for a field, the content will be referred for manual population by a staff member. As these fields are not used to determine whether a claim is approved, this will not delay a cover decision. The Cover Decision Service also uses two statistical models in tandem to calculate the probability of acceptance and case complexity. ACC has used data from 12 million previous, anonymised claims to build its models.

Statistics New Zealand (Stats NZ) uses machine learning to produce provisional estimates of migration in New Zealand. This includes using a unit record machine learning model to classify travellers whose migrant status is uncertain. This is a classification model, applied to individual border crossings. The proportion of travellers in each month who have an uncertain migrant status range from a majority for the most recent few months, to a minority for all other months. This is because most travellers are coming and going within a few weeks and their migrant status can be logically finalised for at least one of their border crossings. The model learns about the features of border crossings that make them more or less likely to be migrant crossings, by looking at millions of historical records and reviewing direction and date of border crossing, the amount of time in orout of New Zealand and time passed since the border crossing. It then applies what it learns to the set of border crossings that are unknown (because not enough time has passed), at an individual crossing level, to

estimate the probability that a particular arrival or departure is a migrant arrival or departure.

The Ministry of Social Development (MSD) uses an algorithm for a service to help find and offers extra help to early school leavers who are not in employment, education, or training (NEET). NEET provides empirical assessment to support a decision that identifies rangatahi who would benefit most from an intervention or policy. The tool helps MSD staff understand the rangatahi's circumstances and make the best referral decisions. The statistical modelling tool is used as just one way to understanding their needs. The individual needs' assessment also helps by providing an opportunity to talk about what else is going on in their lives and how the service may help – including learning to drive a car, access to drug, alcohol or other specialist education courses.

The three following case studies illustrate the utility of XAI spanning three different industries.

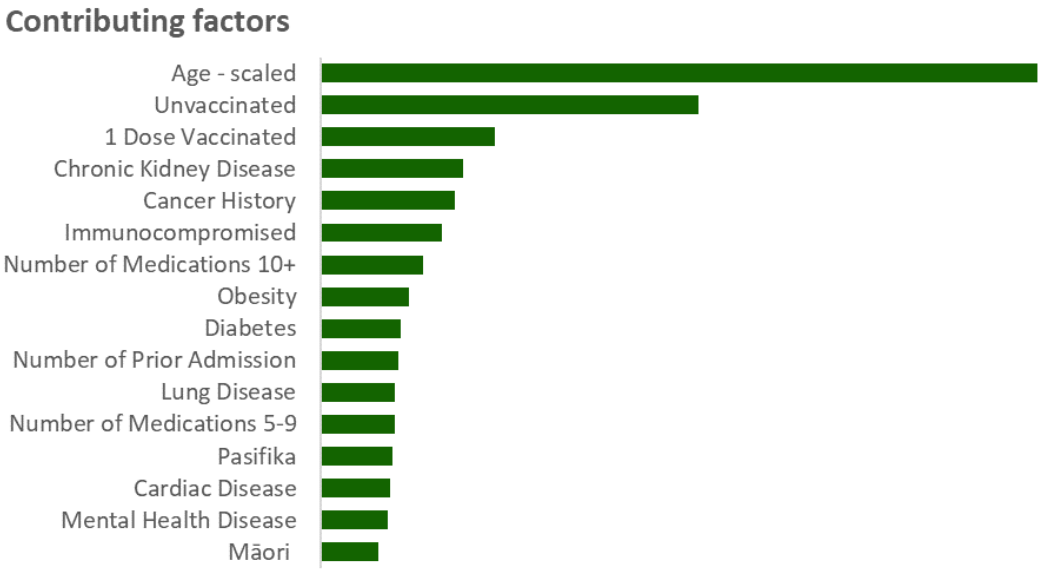
Case Study 1: The interpretable COVID-19 Triage Tool

In 2021, the COVID-19 Delta variant outbreak in Tāmaki Makaurau resulted in adverse health and welfare outcomes that disproportionately affected Māori and Pasifika communities. The impending threat of the Omicron variant exposed limitations with the model of care, where all COVID-19 positive patients were contacted for symptom-driven management. This, combined with weaknesses in existing healthcare tools, which were insufficient in identifying individuals at high risk of clinical deterioration, highlighted the need for an alternative solution.

In response, the Institute for Innovation and Improvement at Te Whatu Ora Waitematā developed the COVID-19 Triage Tool. This tool was tailormade to Aotearoa's specific needs, incorporating data from COVID-19 cases, immunisation records, patient demographics, and comorbidities to predict the risk of hospitalisation for patients with COVID-19 at the point of diagnosis. Notably, it used an XAI logistic regression model. Clinical staff needed a model that was both easy to interpret and quick to understand.

A black box model or a model containing complex feature interactions was out of the question as these could hide spurious relationships. A simple and interpretable model enabled healthcare professionals to easily understand and trust its outputs, interpreting the prediction for a specific patient based on that patient's data. The diagram below shows the importance of the contributing factors to the COVID-19 Triage Tool and

highlights the importance of vaccination on hospital admission rates.



Covid triage model, contributing factors. Please note that age has been scaled in the graph above and represents the risk of the oldest individual (96 years old) when compared to the youngest individuals (18 years old) within the training dataset.

A peer review conducted by Precision Driven Health validated the methodology and accuracy. Data from over 10,652 patients in the early stages of the Omicron outbreak was used to create a second version of the model which was integrated into a visual dashboard for use by health professionals.

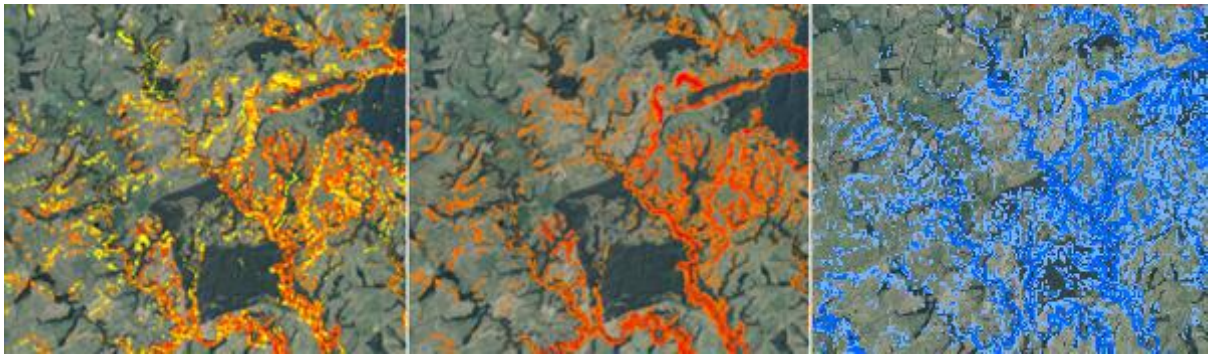
To emphasise understandability, the tool provided clear risk scores and insights into underlying variables. It achieved an AUC ROC of 0.817 (considered very good), particularly excelling in predictions for the Māori and Pasifika communities. By focusing on just 10percent of the population, the model was able to identify approximately 60percent of all admissions and successfully categorised patients into low, medium, and high-risk groups to help target treatment. This triage tool ensured prioritised care through its combination of accuracy and clarity.

The COVID-19 Triage Tool was implemented within COVID-19 Care in the Community Hubs. Across Tāmaki Makaurau, high-risk ethnicities (including Māori and Pacific populations) were managed by ethnicity-specific community care hubs which enabled culturally appropriate care and resources to be funnelled into the high-risk populations. The triage tool ensured prioritised care through its combination of accuracy and clarity. By implementing the model within ethnicity-specific hubs upholding the principles of equity, it demonstrated the potential of interpretable AI in healthcare.

Case study 2: Understanding landslides

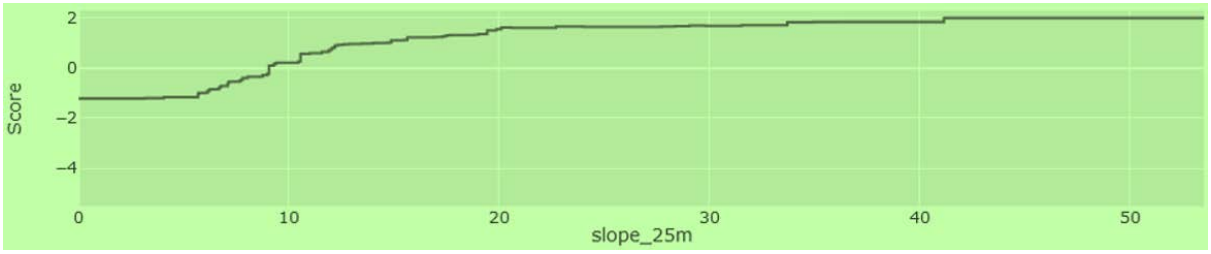
In February 2023, Cyclone Gabrielle led to devastating floods and landslips in the Gisborne and Hawkes Bay regions. The unprecedented climatic events highlight the urgent need for pro-active measures to protect our land and communities. To explore factors contributing to landslips and help identify mitigation strategies, Silver Fern Farms in partnership with environmental data science specialist Lynker Analytics turned to interpretML, an open-source machine learning library that provides glass box models, often performing with similar accuracy to popular ensemble models such as random forest. The glass box model used is fully interpretable, so that not only are feature importance and response for the whole model viewable, but for each prediction made by the model, the exact contribution of the model's inputs can be seen.

The project team assembled data from open sources such as slope, elevation, and land cover alongside previously mapped landslips from Manaaki Whenua Landcare Research. The interpretable model produced a probability map of landslips and achieved 0.83 ROC AUC score when compared to mapped slips. An AUC ROC score gives a measure of accuracy of ranking of predictions. A score of 0.5 suggests random predictions, while 1.0 indicates a perfect prediction. The predictions from the model reveal a strong correspondence between the predicted slips and actual slips as well as a clear relationship with slope and vegetative cover.



(left) slip predictions in yellow. (centre) slip likelihood in orange. (right) high slope in blue.

So far, this is as expected from a probabilistic model. Where glass box models excel is in viewing the response of the model to individual input features such as slope as seen below.



Feature response for slope at 25m scale

A score of zero means that the feature has no effect on the likelihood of slip prediction while a positive number indicates an increase of likelihood, and a negative number indicates a decrease in likelihood. From these we see a strong increase in likelihood of slips with slopes from five degrees to 20 degrees. In summary, the transparent decision-making provided by InterpretML is useful in identifying vulnerable zones and identifying positive landscape changes. As we prepare for future weather events, tools like InterpretML can equip us with the knowledge and foresight to fortify our landscapes and communities.

Case Study 3: An explainable AI model for legal sentencing

The use of XAI in Aotearoa's criminal justice system was recently explored through a proof-of-concept study on assault case sentence prediction. The authors advised against replacing human judgment and instead focused on how to enhance it with data-driven insights. Their study used a simple XAI model, which was trained on 302 New Zealand assault cases. Their results showcased AI's ability to predict sentence length (within a year's accuracy) and crucially, to explain the reasoning behind those predictions. This emphasis on explainability is a step towards ensuring that AI's recommendations are transparent, fostering public understanding and trust in the process.

In Aotearoa New Zealand, the sentencing landscape is complex, balancing individual discretion with the need for systemic consistency. The AI model, with its ability to analyse past decisions, offers a potential tool for judges to align their sentencing approaches. By cross-referencing AI's predictions with past cases and guideline judgments, judges can gain a richer understanding of the various factors influencing sentencing decisions. This could lead to enhanced consistency across the court system, while still maintaining the essential human element in legal proceedings.

Beyond the courtroom, this approach is promising for legal professionals. Lawyers may leverage its predictions to better strategise cases and understand potential sentence outcomes. The explainability may also reveal sentencing patterns, offering insights into how cases should be presented for optimal outcomes. For researchers, the model provides opportunities to critically analyse the justice system, revealing biases and patterns, and initiating discussions on potential improvements.

Appendix

To learn more about explainability:

The origin of XAI (DARPA's Explainable Artificial Intelligence Program): https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/2850

DARPA's explainable AI (XAI) program: A retrospective: https://onlinelibrary.wiley.com/doi/full/10.1002/ail2.61

A foundational XAI method (LIME: Local Interpretable Model-agnostic Explanations): https://github.com/marcotcr/lime

SHAP (SHapley Additive exPlanations): a game theoretic approach to explain the output of any machine learning model:

https://github.com/shap/shap

Explaining Black-Box Machine Learning Predictions - Sameer Singh: https://www.youtube.com/watch?v=TBJqgvXYhfo

Explanation in artificial intelligence: Insights from the social sciences:

https://www.sciencedirect.com/science/article/pii/S0004370218305988

Office of the Privacy Commissioner | Generative Artificial Intelligence – 15 June 2023 update

ICO and The Alan Turing Institute:

Explaining decisions made with AI | ICO

Office of the Privacy Commissioner | Privacy Impact Assessment Toolkit

Home — Web Accessibility Guidance project — NZ Government (govtnz.github.io)

Free living book: https://christophm.github.io/interpretable-ml-book/

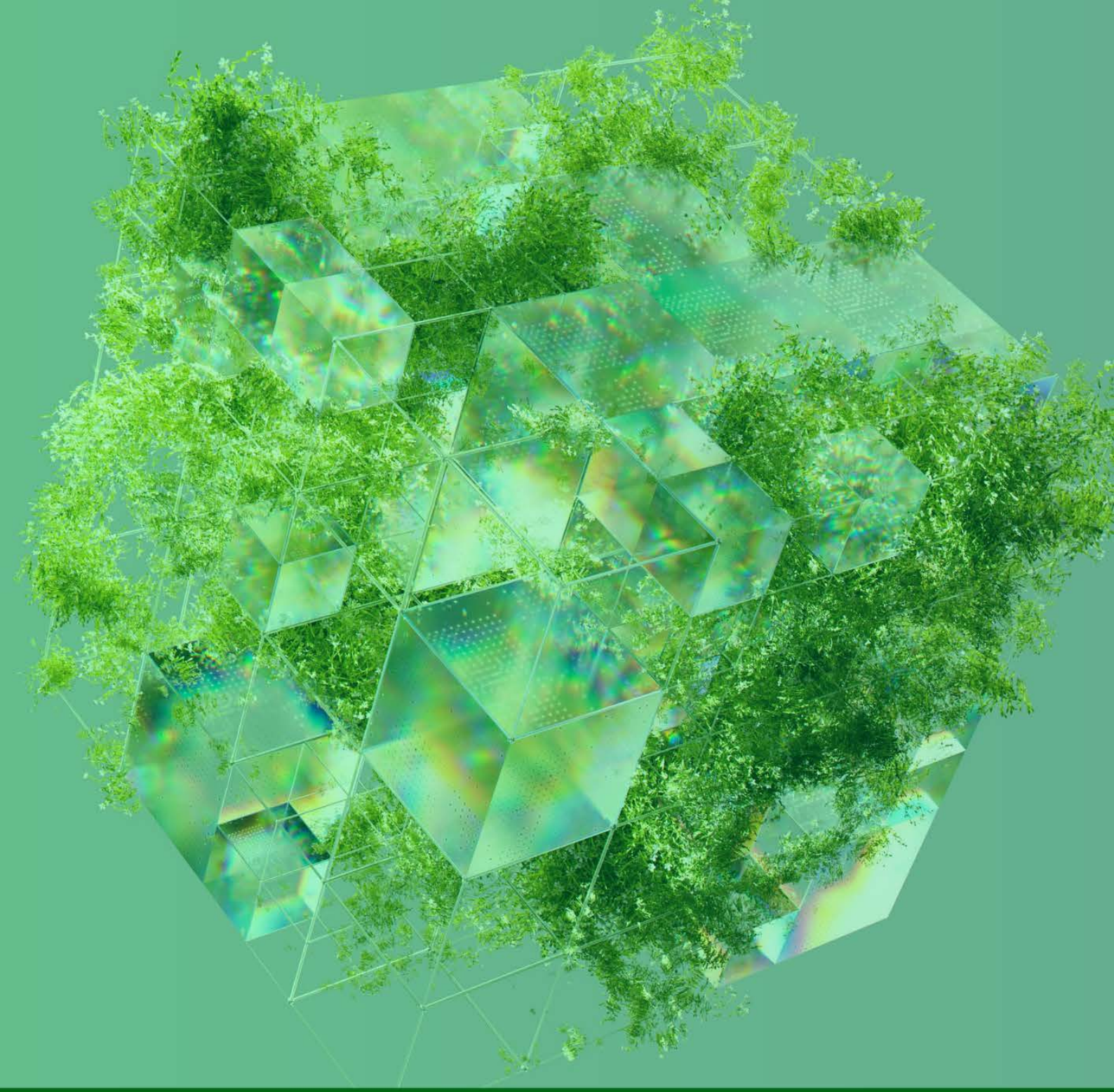
Principles of Māori Data Sovereignty:

https://www.temanararaunga.maori.nz/s/TMR-Maori-Data-Sovereignty-Principles-Oct-2018.pdf

Case study 3:

https://theconversation.com/we-built-an-algorithm-that-predicts-the-length-of-court-sentences-could-ai-play-a-role-in-the-justice-system-193300

Te Tiriti o Waitangi Principles:
https://www.waitangitribunal.govt.nz/assets/Documents/Publications/WT-Principles-of-the-Treaty-of-Waitangi-as-expressed-by-the-Courts-and-the-Waitangi-Tribunal.pdf



AI Forum New Zealand
Te Kāhui Atamai Iahiko o Aotearoa

Explainable AI – building trust through understanding

November 2023

info@aiforum.org.nz
aiforum.org.nz
+64 9 394 7693

PO Box 65503
Mairangi Bay
Northshore 0754

A discussion on explainable artificial intelligence (XAI) and how it can be used to enhance trust in AI systems developed in Aotearoa New Zealand.