# Principle 2: Human-centred values

Let's see how this principle guides responsible AI practice.  We'll look at a key risk this principle is designed to address and explore a brief practical example of how this principle relates to an AI use case.

**Definition:** Public Service AI use should respect the rule of law, democratic values and human rights and labour rights through the lifecycle of each AI system or product.

These rights and laws include personal data protection and privacy, dignity, non-discrimination and equality, self-determination and autonomy. Public service workers have the right to be consulted on changes made to their work and working arrangements. Agencies need to provide human oversight throughout the AI lifecycle to ensure ethical and appropriate use.”

**Let’s focus on: Bias:** AI systems can perpetuate and amplify existing biases in society, leading to unfair or discriminatory outcomes. AI works by predicting and recreating patterns, not creating a “correct outcome”. When the training data is skewed or incomplete, bias can result. For example, without the input of disabled people, AI systems might ignore their needs or suggest irrelevant or incorrect solutions for the community.

**Practical example:** Government agencies across New Zealand are looking to fill roles in the public service. Selecting the right candidates can be time-consuming, and AI could potentially streamline this process. However, imagine these agencies deploy an AI system that inadvertently discriminates against women and people with disabilities. Why?

First, the AI might have been trained on historical hiring data favouring non-disabled male candidates — data reflecting past human biases that weren't corrected before use. Consequently, the AI learned to replicate these biases. Also, decision-makers might trust the AI's recommendations without realising its flaws, thus continuing biased hiring practices.

Moreover, AI models often draw from datasets predominantly sourced from Western contexts, which may carry societal biases. The complexity of these models makes it hard to pinpoint where bias occurs, and once biased outputs influence media or reports, they reinforce harmful stereotypes. It is essential for us to recognise these risks and actively work to minimise bias in AI-driven processes.

## Spotlight: The Role of Data in AI

Data is the fuel that drives how AI systems learn, make predictions, generate content, and augment decisions. In the field of AI, you may hear people say "garbage in, garbage out" to highlight that the outputs of AI systems reflect the quality of their inputs. The importance of data is relevant to all types of AI: from machine learning models predicting service demands, computer vision systems analysing infrastructure, to generative AI assisting with communications.

Throughout its life cycle, an AI system interacts with various data types. Useful questions to consider when planning or assessing an AI use case are:

1. **Training: What data will/did the AI learn from?** Sometimes we will train a model ourselves (which needs data to learn). We need to use personal information in ways that fit the purpose for which it was collected.

**Specialisation: What data will we add to a pre-trained model to specialise it?**We often use pre-trained general-purpose models instead of training a model from scratch.  Is the training data behind an AI system relevant, reliable, and ethical?

1. **Usage: What data is fed into the system during use, and what is created?**We commonly provide data to an AI system and create new data from it. Information (such as prompts) that go into an AI system need to be kept safe.

Good data stewardship involves ensuring data is accurate, fair, lawful, representative of our communities, and securely handled at all times.

## Types to AI Lifecycle Stages Related to an AI Chatbot Use Case

Scenario: Imagine an agency has developed an AI-driven chatbot designed to handle internal queries, efficiently assisting our people with their questions. The chatbot relies on various data types throughout its lifecycle to function effectively. · These data types match the appropriate stage in the AI lifecycle.

**AI Lifecycle Stage:** Training: What data will/did the AI learn from?

**Data Type:** Public internet text, such as Wikipedia, books, articles, and open-source research. This data is used to create the foundational large language model that powers the AI chatbot's understanding of language

**AI Lifecycle Stage:** Specialisation: What data will we add to a pre-trained model to specialise it?

**Data Type:** Internal knowledge base and organisation FAQs. These resources are used for retrieval-augmented generation, allowing the chatbot to provide accurate and relevant answers to specific internal queries.

**AI Lifecycle Stage**: Usage: What data is fed into the system during use, and what is created?

**Data Type:** Interaction data, such as questions like "How do I reset my password?" and logs of chat interactions. Additionally, analytics on chat usage help refine the chatbot's performance by analysing how employees engage with it.

## But are you allowed to use that data?

AI techniques often rely on ingesting data to learn, specialise and engage in everyday interactions, as outlined above. However, this extensive data collection can sometimes conflict with regulatory principles, particularly those of the Privacy Act. These principles state that personal information should be:

1. Limited to what is necessary,
2. Collected by legal and fair means, and
3. Gathered with the individual’s consent or knowledge.

Here are some examples of key privacy risk questions to consider when evaluating and planning AI use cases:

Will the AI access only the essential data needed for its function?

* Concerning scenario example: An AI service for HR efficiency requires minimal data about skills and job titles but unintentionally includes names and salaries, leading to a privacy breach.
* Risk mitigation: Verify how data governance and hygiene are maintained: Ensure that data used in AI models is limited to essential information and well-managed to reduce risk exposure. Implement practices such as sensitivity labelling and access management to support data loss prevention.

Are you aware of all parties involved in providing the AI?

* Concerning scenario example: An AI service meant to handle customer complaints retains customer details, which are then sold by the vendor to third parties. This results in a breach of the New Zealand Privacy Act and potential fines.
* Risk mitigation: To fully manage potential risks, we need to understand the role of all parties, including any vendors or third-party providers that may handle our data. Open-source models can present unique security and compliance challenges. Conduct thorough evaluations to ensure they meet necessary standards and mitigate associated risks.

So what: Key takeaways for evaluating use cases.

**Does the use case...**

* Respect human and labour rights.
* Provide human oversight to ensure ethical and appropriate AI use?
* Identify and address potential biases?
* Evaluate the impact on vulnerable communities?
* Avoid perpetuating historical biases?
* Limit AI data access to essential information only so as to maintain privacy?
* Understand all parties involved in AI provision to manage security and compliance risks?