# Principle 4: Safety and security

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 systems should treat the security of customers and staff as a core business requirement, not just a technical feature (security-by-design). They should minimise risk to individual or national safety and security under normal use, misuse or adverse conditions.

The Public Service should ensure traceability of data, apply a robust risk management approach and work collaboratively with commercial and security colleagues in the procurement and assurance of AI tools.”

**Let’s focus on: Security:** Because of the large amount of data that AI systems leverage and generate, they have become a target for data theft and are also at risk of large unintended data breaches. As AI becomes more deeply integrated into core business processes with agentic AI, it becomes increasingly important to safely handle sensitive data that exists across those processes.

**Practical example:** A Generative AI chatbot is made directly available to the public to answer questions about the treatment and financial support that is available to them after having an accident. To protect privacy, any personal information disclosed by the user is not stored. Some privacy settings are controlled by a system level prompt, which cannot be accessed by front-end users – with safeguards in place to deter bad actors who intentionally try to bypass this.

When asked questions about how personal information is used, the chatbot explains that such information is redacted / not stored – but stops short of sharing detail on its system settings that would be harmful in the wrong hands. This helps the user understand the mechanism of protection, to foster greater trust in the system they are using.

## Spotlight: Security of AI Systems

[The Responsible AI Guidance for the Public Service](https://dns.govt.nz/standards-and-guidance/technology-and-architecture/artificial-intelligence/responsible-ai-guidance-for-the-public-service-genai/genai-foundations/security) updated this year highlights that security plays a vital role in enabling the reliability and resilience of GenAI systems. Like all digital systems, GenAI systems can be susceptible to security vulnerabilities and misconfigurations.

Security breaches can harm our reputation, negatively impact customers, compromise privacy, and erode trust in public services. We need to store data in systems aligned with security practices to ensure adequate protection.

AI systems should be robust, secure and safe throughout their entire lifecycle. Organisations should look to implement robust assessment processes and controls to identify and manage the risks relevant to their use of AI. Assessments should capture relevant business and technical details and may require involvement from a range of business areas such as Legal, Privacy, Security and others. Here are some examples of key security risk questions to consider when evaluating and planning AI use cases:

**Where will your AI model be hosted and how will it be integrated into your systems?**

* Concerning scenario example: A vendor offers a new integration to a popular AI model that enhances an existing solution in your business. However, to use this feature the model transfers your organisation's data to an offshore provider via an API, lacking clear security controls and transparency, which poses a risk of data exposure.
* Risk mitigation: Before you start, it is vital you conduct a risk assessment including considering the data you have, and your jurisdictional risk responsibilities. Advice about potential jurisdictional risks can be found on DIGITAL.GOVT.NZ. You should discuss these concerns with your supplier. Find out how your suppliers provide access to the AI models and where your data might be stored and shared. It is essential to document these connections and understand where underlying AI services are sourced from and hosted, to be able to fully assess potential risks. Many models might not be directly integrated, requiring attention to how data is transferred. This is vital for understanding exposure, especially with models not natively integrated into your systems.

**Will you have visibility of AI system usage?**

* Concerning scenario example: Without proper oversight, attempts to access restricted information might go unnoticed. Cybercriminals could exploit and attack AI services used by your staff or customers, risking unauthorised access to your information.
* Risk mitigation: We will need to track interactions, identify threats, and ensure proper functioning of AI models we are implementing. Develop processes to understand acceptable user interactions with your systems. Based on these patterns we can then detect risky and unusual behaviour. Ensure all AI systems used by your agency are certified and accredited before they are made available to users, as advised within the New Zealand Information Security Manual (NZISM) and your security teams.

**Are existing enterprise controls applied to our new AI services?**

* Concerning scenario example: A vendor proposes a new AI service to be used with an existing system which you host in your own cloud tenancy. This AI model can be deployed in various ways, which may give you different levels of oversight and control.
* Risk mitigation: We should attempt to leverage existing enterprise security controls and processes, maximising existing security investments. Ensuring we are following a Secure By Design (SBD) methodology to identify risks early in the design process, and plan how existing controls such as vulnerability and patch management, monitoring and alerting or identity and access management can apply. Choosing to consume services in a way that offers better visibility (such as hosting services in your cloud tenancy) may provide stronger controls.

**Are you planning for technology advancements and adjusting controls accordingly?**

* Concerning scenario example: Existing controls may not remain capable of managing risks as AI technology evolves. The recent release of Agentic AI systems challenges common control assumptions highlights how existing controls could become ineffective as this technology evolves.
* Risk mitigation: Implementing ongoing risk reviews which re-baseline control effectiveness, and remaining aware of impactful incoming technology changes can support you to manage risk as it evolves. Ongoing investment into controls, people and processes will likely be required in the short and medium term due to the exponential growth and uptake of AI services.

**Resources:**

* [Security and GenAI](https://dns.govt.nz/standards-and-guidance/technology-and-architecture/artificial-intelligence/responsible-ai-guidance-for-the-public-service-genai/genai-foundations/security#resources)
* [New Zealand Information Security Manual (NZISM)](https://nzism.gcsb.govt.nz/)
* [Cloud Jurisdictional Risk guidance](https://www.digital.govt.nz/standards-and-guidance/technology-and-architecture/cloud-services/assess-the-risks/cloud-jurisdictional-risk-guidance). More detailed, country-specific guidance is available on request to GCDO.
* [Joint Guidance:](https://www.ncsc.govt.nz/news/deploying-ai-systems-securely)
	+ [Engaging with Artificial Intelligence](https://www.ncsc.govt.nz/news/engaging-with-ai)
	+ [Guidelines for Secure AI System Development](https://www.ncsc.govt.nz/news/secure-ai-system-development)
	+ [Deploying AI Systems Securely](https://www.ncsc.govt.nz/news/deploying-ai-systems-securely)

Decide if this statement is True or False

**Understanding where AI models are hosted and how they connect to your systems is essential for managing security risks**

 Correct Answer: “True” · Feedback if correct: Correct. Knowing where AI models are hosted and how they are integrated allows you to assess risks like data exposure, weak connections, or supplier vulnerabilities. Clear documentation of hosting and data flows is a key part of risk management

Feedback as to why it is not “False”: Hosting and integrating details are critical for assessing security risks. Without a clear understanding of how and where your data travels, it can be difficult to spot potential vulnerabilities or manage supplier risks effectively.

So what: Key takeaways for evaluating use cases.

**Does the use case... ·**

* Treat security as a core business requirement. Ensuring data protection and privacy
* Minimise risk of data theft and unintended breaches through robust risk assessments and security measures
* Document connections to understand where AI services are sourced and hosted
* Provide visibility into AI system usage to track interactions and identify threats
* Plan for technological advancements, continuously adjusting controls to remain effective?