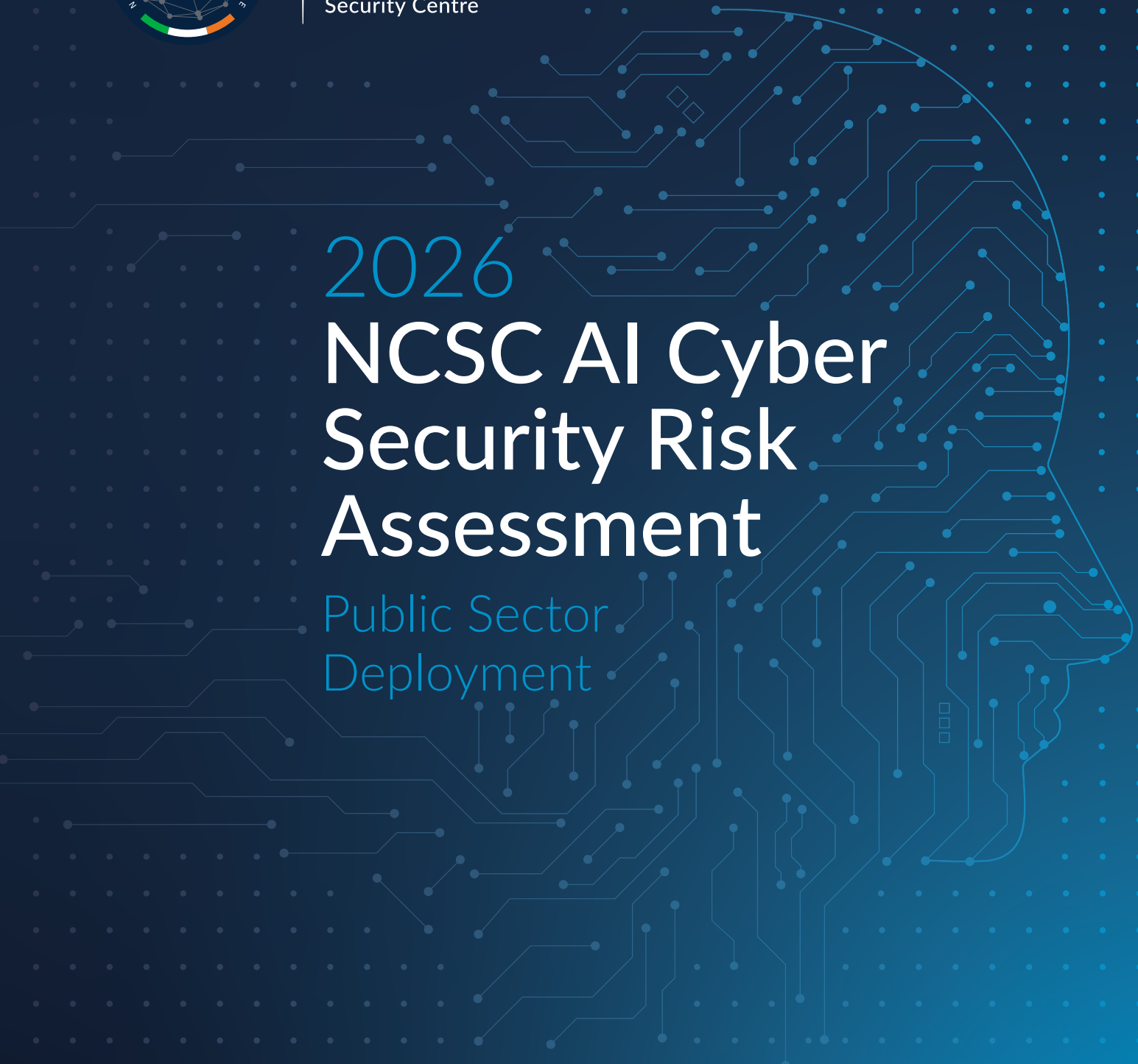
An Lárionad Náisiúnta Cibearshlándála
National Cyber Security Centre

# 2026 NCSC AI Cyber Security Risk Assessment

## Public Sector Deployment



www.ncsc.gov.ie

Rialtas na hÉireann
Government of Ireland

> The secure deployment of AI can transform the delivery of public services through increased operational efficiency and effectiveness.

# Executive Summary

The secure and safe deployment of AI by public sector bodies can significantly enhance operational efficiency, improve service delivery, and enable data-driven decision-making in the public sector. NCSC-IE is fully supportive of AI adoption across the public sector and has produced this assessment to help public sector bodies manage the cyber security risks that come with it.

Europe and Ireland's regulatory framework sets clear obligations. The EU AI Act, NIS2, GDPR, and several national legal and policy frameworks collectively establish requirements that public sector bodies must meet when deploying AI.

AI fundamentally changes an organisation's risk profile. It expands the attack surface, increases data movement across systems, and introduces behaviour that can change over time as models are updated and interact with new data. Unlike traditional systems, new vulnerabilities can emerge after deployment, making cyber security for AI a continuous process, not a one-time exercise.

The threat landscape regarding AI integration is complex and constantly evolving, comprising both state aligned and non-state threat actors. A broad understanding of both their motives and capabilities is vital for public sector organisations when assessing risks and developing mitigation strategies. It is important to be cognisant of insider threats as well as external attacks when considering protective measures.

Security must be embedded at every phase of the AI lifecycle, from design through to end-of-life. A compromise at any stage can affect the integrity of the entire system. Risks identified during design and development must be resolved before deployment. Risks that emerge post-deployment require continuous monitoring and response. Systems approaching decommissioning must be secured with the same rigour as live production systems.

Key assets at risk include sensitive and classified information, citizen data, identity and access management infrastructure, cyber security and network infrastructure, AI models and agents, vendor and supply chain relationships, and the operational processes underpinning service delivery. Agentic AI systems, which operate with elevated credentials and can execute actions at scale, represent an emerging and specific area of risk that organisations actively exploring this technology should prioritise.

Public sector bodies should treat this assessment as a baseline for assessing their AI cyber security risk posture. Immediate priorities are establishing governance and usage policies for AI (including for tools already used informally) and identifying key assets at risk. The comprehensive guidelines published alongside this assessment provides specific, lifecycle-mapped controls to support that work.

# Contents


Executive Summary 3

Introduction 5

Scope 6

Methodology 7

Summary of Findings 8

Regulatory Landscape 10

> Legislation 10

> Strategy & Frameworks 11

Threat Landscape 12

> Threat Actors 12

> Assets at Risk 13

> Primary Threat Types 16

Risk Scenarios across the AI Lifecycle 22

1. Design 23

2. Development 24

3. Deployment 25

4. Maintenance 27

5. End-of-Life 28

Summary of Risks 30

Recommendations & Next Steps 31

Annex I 33

> Methodology 33

Annex II 34

> Survey Results 34

Annex III 38

> Glossary of Key Terms 38

Annex IV 40

> Bibliography 40

# Introduction

The secure deployment of AI can transform the delivery of public services through increased operational efficiency and effectiveness. The National Digital and AI Strategy[1] seeks to deliver on this potential through accelerated transformational change in the public sector with AI adoption as a key focus. This AI risk assessment has been developed by the National Cyber Security Centre (NCSC-IE) as an enabler in supporting this transition.

The possibilities and benefits offered by embracing AI in the public sector are vast and varied. Implementing AI safely can offer significant improvements, from unlocking innovative ways of working to automating complex, repetitive tasks and enhancing decision-making through advanced analytics.

The National Cyber Risk Assessment[2] identified evolving technology, specifically AI, as a systemic risk, necessitating a secure and proactive approach to its integration by public sector bodies. While governance, trust and capability challenges may dominate organisational considerations prior to implementation, it is also important to acknowledge that adopting AI solutions can alter the cyber security, data, and privacy risk profile, increasing the attack surface area, system complexity and impacts associated with exposure of sensitive data.

The risk environment is more complex, and dynamic compared to more traditional systems. The behaviour of AI systems can change over time due to new data, model updates, and evolving interactions with users and environments. As a result, new vulnerabilities may emerge, even after deployment. This makes cyber security for AI an ongoing process, requiring continuous monitoring, reassessment, and adaptation of controls throughout the entire lifecycle.

While NCSC-IE acknowledges the risks posed, it also recognises the enhancements and the positive impacts offered by the safe and secure implementation of AI and is fully supportive of its adoption across the public sector.

*To support public sector bodies* with the safe deployment of AI, NCSC-IE are:*

1 **Assessing the cyber security risks** associated with the integration of AI

2 **Raising awareness** among key stakeholders

3 **Providing guidance** to enable the secure deployment of AI

This document is designed to address the first two points and will inform a revision of the cyber security guidance on AI for public sector bodies which should be read in conjunction with this risk assessment.

**Refers to organisations and entities established, funded, or controlled by the Government to provide services or implement policies that serve the public interest.*

# Scope

AI deployment refers to the process of integrating a trained artificial intelligence model, including large language models (LLMs) into a real-world environment. This integration enables the AI to make decisions, automate tasks, or generate insights, thereby enhancing operational efficiency and effectiveness across various sectors.

**Who this risk assessment is for:** This assessment highlights the most significant cyber security risks associated with the deployment of AI across the public sector. Although specifically intended for Chief Information Officers, Chief Information Security Officers, Chief Technical Officers and senior managers responsible for the delivery and management of ICT services and systems across government departments. This risk assessment may also be useful for management boards who will have expanded responsibilities under NIS2. It is also applicable to other sectors who have or plan to integrate AI in their organisations.

**How to use it:** This risk assessment provides a starting point to understand the current threat landscape. While it does not cover all risks, the assessment and the subsequent NCSC-IE guidelines should instil confidence in organisations enabling informed cyber security decisions regarding AI integration, allowing them to leverage AI capabilities securely to deliver public services more efficiently.

# Methodology

*A five-phased approach was used in the development of this risk assessment:*

This process enabled NCSC-IE to understand current and planned AI deployments across public sector organisations in the context of the current threat environment. It was informed by global cyber security attacks on AI deployments which were also used in the development of risk scenarios and the presentation of relevant case studies.

A detailed outline of the methodology applied is available in Annex I.

# Summary of Findings

The results from NCSC-IE's survey and interview programme indicates that AI adoption across the Irish public sector is at an early but accelerating stage. Most organisations have some policies and basic security controls in place, and the majority are actively planning to expand their AI use over the next twelve months, with Agentic AI emerging as a near-term consideration for a significant number of respondents.

Against this backdrop of growing adoption, several findings warrant specific attention.

### Maturity levels remain low

While basic governance foundations exist in most responding organisations, the depth of AI-specific security capability is limited. Few organisations have AI-specific monitoring in place, and access to staff with the necessary skills to manage AI security risks was consistently identified as a challenge.

### Access to AI tools is uneven

In half of responding organisations, cloud-based LLMs are available only to a subset of users. This reflects both a cautious and proportionate approach in some cases, and in others a gap between sanctioned and unsanctioned use that increases shadow AI exposure.

### Governance concern for organisations

Respondents identified ungoverned use of AI tools as a current concern. This finding reflects a gap between the pace of AI adoption by individual staff and the pace at which governance frameworks are being established.

### Data quality and data protection are the primary operational challenges

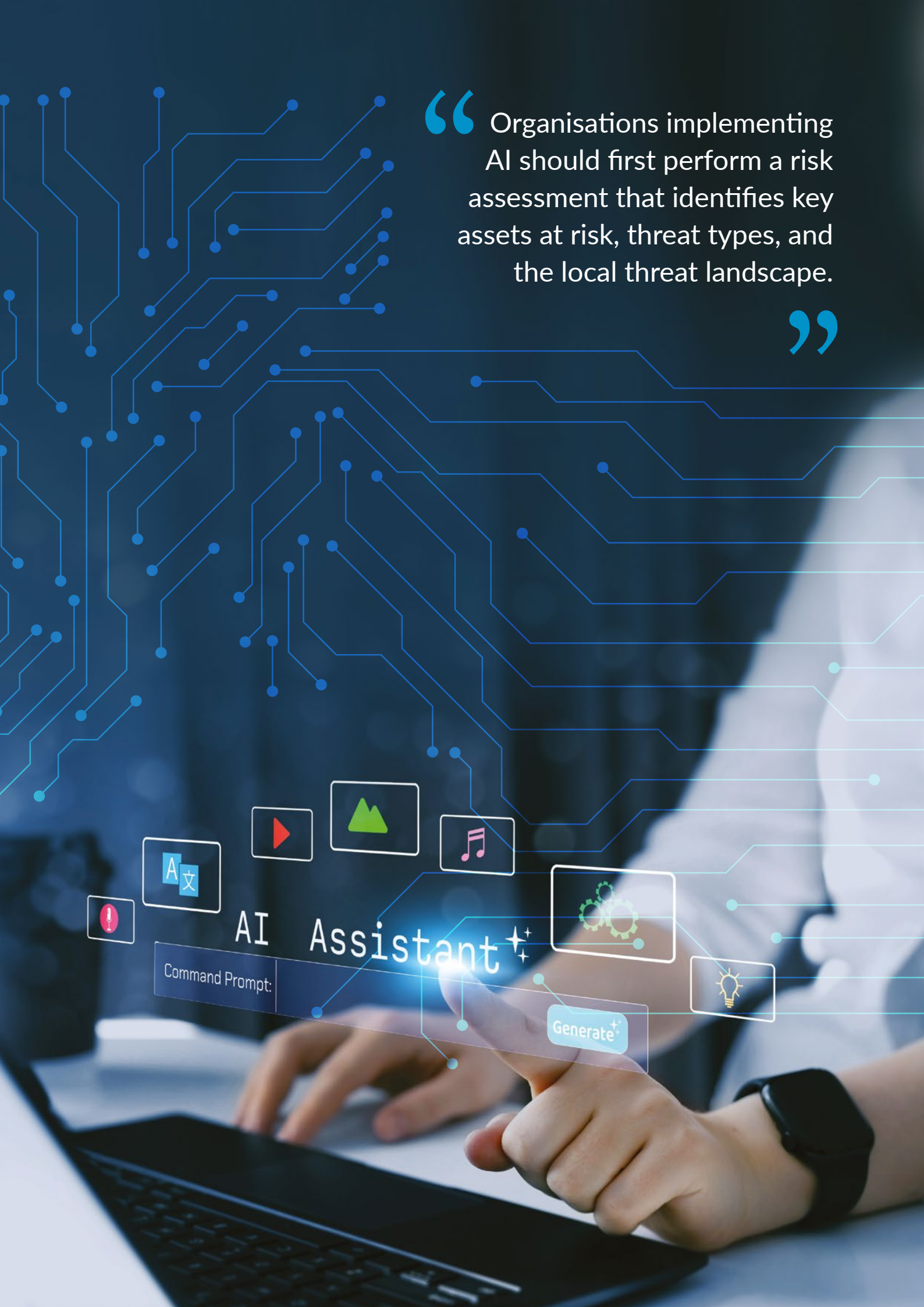
Respondents consistently identified these as the most significant barriers to secure AI implementation, ahead of technical infrastructure concerns. This points to a need for guidance that addresses data governance alongside cyber security risk management.

### Organisations are looking to NCSC-IE for practical support

Respondents identified guidance on secure AI implementation, procurement support, training frameworks, and a checklist of baseline security controls as their most valued areas of assistance. This assessment and the accompanying guidelines are a direct response to that need.

Given the early stage of adoption, organisations implementing AI should first perform a risk assessment that identifies key assets at risk, threat types, and the local threat landscape. This baseline is essential for designing mitigations and realising the benefits AI offers to service delivery in a safe and secure manner.

> Organisations implementing AI should first perform a risk assessment that identifies key assets at risk, threat types, and the local threat landscape.



# Regulatory Landscape

## Legislation

The EU AI Act[3] is foundational in establishing Irelands regulatory landscape. It categorises AI systems based on their risk levels and delineates requirements for development, deployment, and monitoring to ensure ethical practices and user safety. The Act prioritises transparency, requiring organisations to disclose how their AI systems function, thus fostering public trust. It mandates that high-risk AI systems shall be designed and developed to achieve an appropriate level of cyber security throughout their lifecycle and be resilient against attempts by unauthorised third parties to alter their use, outputs, or performance by exploiting system vulnerabilities.

To further bolster cyber security, the NIS2 Directive[4], which will be transposed by the upcoming National Cyber Security Act, imposes obligations on organisations to assess risks, implement security measures, and ensure incident reporting transparency across essential services. NIS2 is particularly relevant for AI systems, as defined in the EU AI Act, deployed in public and private bodies in critical sectors including healthcare, energy, and finance, where vulnerabilities can pose significant threats to public safety and service continuity.

Compliance with the General Data Protection Regulation (GDPR)[5] is a critical consideration when implementing AI, particularly as public sector bodies handle significant amounts of personal data. The GDPR mandates strict data protection measures, consent management, and individual rights, ensuring that personal information is processed lawfully and transparently, thereby mitigating risks related to privacy violations.

# Strategy & Frameworks

Irelands National Digital and AI Strategy[1] aims to harness the potential of AI technologies while addressing economic and social challenges. It encourages innovation, promoting an ecosystem where public sector bodies can effectively integrate AI to enhance efficiency and public engagement.

The Department of Public Expenditure, Infrastructure, Public Service Reform and Digitalisation (DPER)[6] provides essential guidelines tailored to assist public sector bodies in navigating the complexities of AI implementation. These guidelines emphasise best practices in governance, accountability, and compliance, ensuring that organisations align with national policies and international norms.

Additionally, the European Union Agency for Cybersecurity (ENISA)[7, 8,] offers frameworks and resources to guide organisations implementing security measures for AI systems, helping to address the unique cyber security challenges posed by AI technologies. The National Institute of Standards and Technology (NIST) AI Risk Management Framework[9] further serves as a vital resource, providing guidelines for managing risks associated with AI systems, emphasising the importance of a secure and trustworthy AI ecosystem. The MITRE Adversarial Threat Landscape for Artificial-Intelligence Systems (ATLAS)[10] framework which maps adversary tactics, techniques and procedures targeting AI and ML systems can be used to help identify threats and test AI resilience. Furthermore, OWASP's Top 10 for Large Language Models[11] and for Agentic Applications[12], outline common AI security risks with recommended mitigation which can help to prioritise and secure AI development.

To meet cyber security obligations the NCSC-IE recommends adopting the **Cyber Fundamentals (CyFun) Framework**[13] which is a recognised, structured tool to assist entities in meeting their NIS2 obligations. CyFun provides a tiered, standards-based framework grounded in the US NIST Cybersecurity Framework v2.0. Certification against this framework is expected to be available in Ireland in 2027. Many of the controls outlined in CyFun are applicable to security of AI systems across the whole AI lifecycle.

This comprehensive regulatory framework outlines the standards and obligations necessary to effectively manage the cyber security risks associated with AI deployment in the public sector, ensuring that AI technologies are used responsibly and securely.

# Threat Landscape

The threat landscape for public sector bodies deploying AI is complex and continuously evolving. Understanding the fundamental components, such as threat actors and vulnerabilities across AI systems is essential for developing effective risk assessment and mitigation strategies.

## Threat Actors

Threat actors can be categorised into state and non-state groups:

**State aligned actors** and their proxies typically tend to be well-funded and resourced, possessing the skills to conduct advanced malicious cyber activities to further both tactical and strategic objectives.

**Non-state actors** (e.g. cybercriminals and hacktivists) are often motivated to conduct cyberattacks based on financial incentives or ideological factors. While insider threats traditionally involve malicious intent, they can also stem from the unauthorised use of AI technologies.

Both state aligned and non-state threat actors can exploit vulnerabilities in AI systems deployed in the public sector, compromising the integrity, confidentiality, and availability of systems to achieve their objectives.

# Assets at Risk

When adopting and implementing AI within the public sector, several key assets are at risk. Identifying assets which are critical to the functioning of public sector bodies is vital for assessing risk and developing mitigation strategies. While each organisation must assess their own asset inventory, it should consider the following:

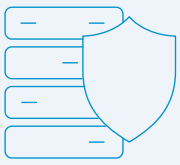
## Data

Personally Identifiable Information (PII), classified information from other organisations (e.g. EU and NATO), and confidential data such as intellectual property, commercially or legally sensitive information, and national security data are prime targets for unauthorised access, theft, or misuse. Poorly implemented AI systems may inadvertently expose sensitive data to unauthorised users; overly permissive open-access levels and uncontrolled sharing of documents and training artefacts increase this risk. Such data breaches could lead to significant violations of privacy and erode trust in government institutions. Furthermore, training data used in AI development often includes sensitive information, which if not carefully curated, can also become vulnerable, exacerbating the risks of exposure during AI operations.

## Identity and Access Management

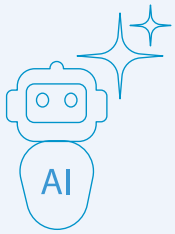
Policies and technologies that determine who and what can access specific resources and what actions they can perform are essential for safeguarding sensitive data and maintaining system integrity. Poor implementation can lead to unauthorised access to critical assets, data breaches and operational disruptions that affect service delivery and public trust. Ineffective AI deployment can lead to flaws in access control protocols, for example poor control of data transfers between the AI system and data sources or inputs, allowing unauthorised users to exploit these vulnerabilities.

## ICT and Cyber Security Infrastructure (including AI models)



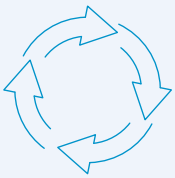
The networks, servers, supporting hardware and software, and the AI models themselves are critical components that can be targeted by cyberattacks, leading to operational disruption and loss of service availability. Poorly hosted, configured and integrated AI systems or models can introduce vulnerabilities like weak network configurations, insecure software interfaces, exposed model APIs, or unprotected model weights, making the infrastructure an attractive target for attacks. For example, unpatched software or model-serving vulnerabilities can facilitate unauthorised access, compromising sensitive data or training sets and critical applications.

## AI Agents



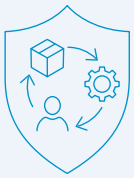

AI agents are critical assets that execute and automate tasks on behalf of the AI system by receiving model outputs or instructions, translating them into actionable operations. Agents often hold elevated credentials and connect to databases, APIs, and operational tooling. Poorly integrated models put agents at risk by increasing their exposure and the likelihood they will be misused or compromised. Compromised agents can execute unauthorised actions, move laterally, exfiltrate data and cause widespread disruption.

## Operational Processes

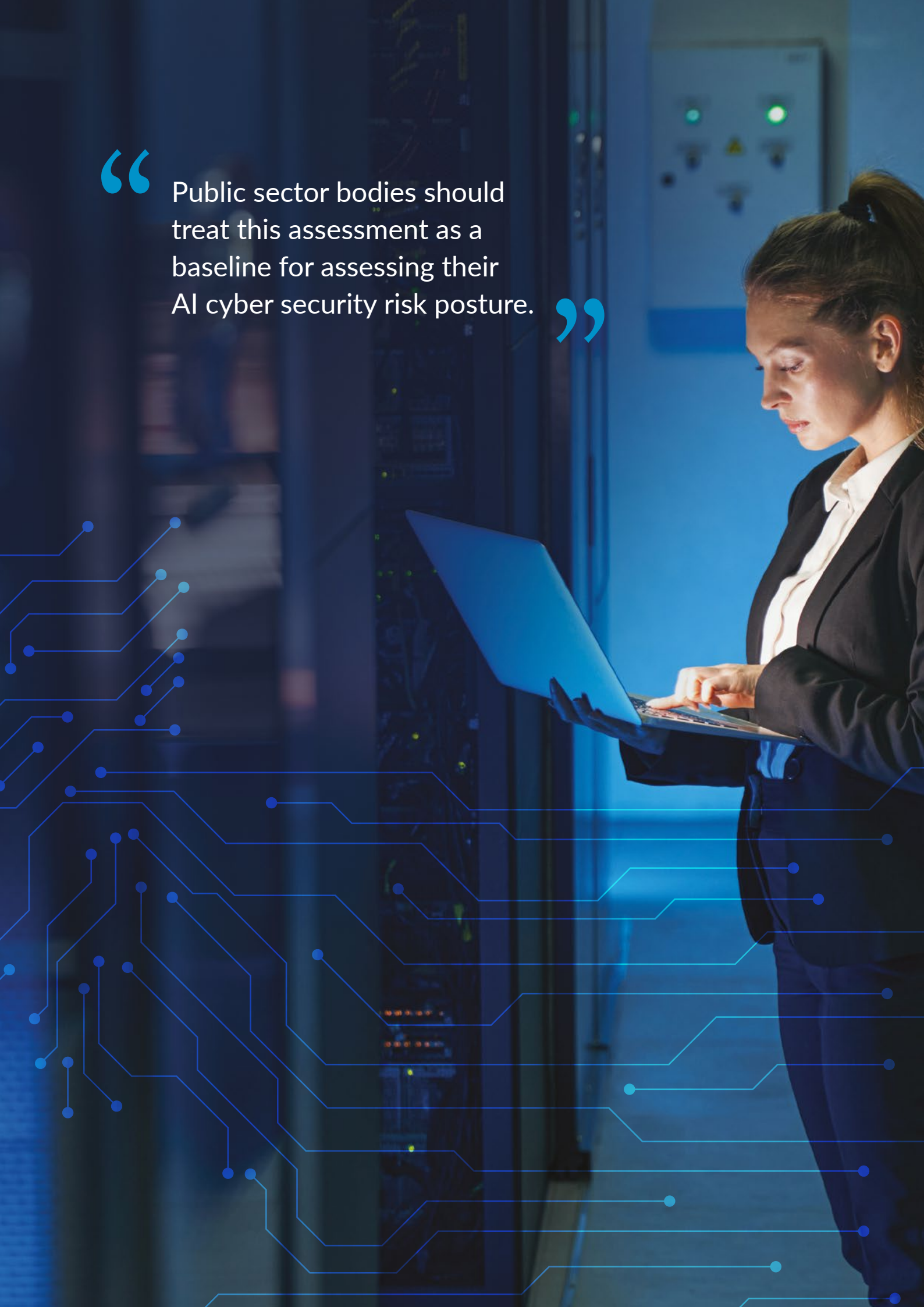


The processes and workflows that govern public sector bodies exist to ensure efficient functionality, if these processes are compromised it can lead to significant systemic failures affecting public services. AI systems, when designed poorly, can result in disruption to workflows by producing incorrect outputs or generating delays in data processing. For example, if an AI misinterprets data, it can result in erroneous decision-making, leading to inconsistent service delivery or even the withdrawal of essential services during critical times. Such disruptions can diminish public trust and hinder overall service efficacy.

## Vendors



Vendor and supply chain relationships are essential assets for public sector bodies. When adopting and implementing AI technologies, these partnerships can facilitate access to essential tools, data and expertise that enhance operational efficiency. However, they also introduce significant vulnerabilities. If a vendor fails to comply with data security protocols or if their systems are compromised, sensitive data handled by AI systems may be jeopardised, leading to data breaches that disrupt service delivery. For instance, if an AI's real-time data provider is compromised, the resulting breach could disrupt several public services simultaneously, undermining their effectiveness. Additionally, poor vendor management can lead to integration difficulties, which may result in vulnerabilities that compromise the overall security posture of the organisation. This could severely impact confidence in service reliability, as users expect seamless and secure interactions with public sector bodies.



> Public sector bodies should treat this assessment as a baseline for assessing their AI cyber security risk posture.

# Primary Threat Types

This section outlines the primary threat types associated with AI deployment in public sector bodies. While not an exhaustive list, it highlights some of the key threat types[11,12,] organisations should consider in their evaluations.

## Prompt injection

Prompt injection occurs when malicious actors manipulate inputs to alter the intended behaviour or output of an AI model. The fundamental vulnerability in prompt injection is that AI models cannot reliably distinguish between legitimate instructions and malicious content embedded in the data they process.

There are two types of prompt injection to consider:

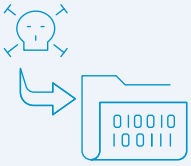
**Direct prompt injection** occurs when a malicious actor's input directly alters the model's behaviour or output.

**Indirect prompt injection** occurs when the model's behaviour is manipulated due to malicious input embedded in external sources, such as websites, files or emails accessed by the model.

Successful prompt injection can lead to various risks, including data exfiltration as well as leaking information about the AI system's infrastructure and prompts. Additionally, it can result in the model generating biased or incorrect outputs because of manipulated content and may compromise critical decision-making processes.

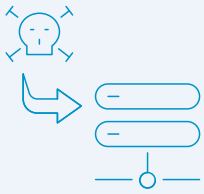
Furthermore, it can enable the execution of arbitrary commands in connected systems, facilitating lateral movement within enterprise environments[14]. It can also cause goal hijacking, undermining intended safeguards and steering automated behaviours towards malicious outcomes. Such risks underscore the necessity of safeguarding against input manipulation to ensure the integrity of the model.

## Data Poisoning



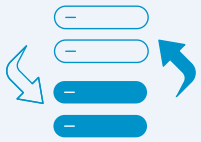

Data poisoning occurs when adversaries introduce malicious data into the training dataset by altering labels or introducing data samples designed to confuse the model. This manipulation causes the model to learn from incorrect information and patterns, resulting in poor performance or misclassification of legitimate inputs. Organisations can mitigate against these risks by implementing robust validation processes and monitoring mechanisms to maintain data integrity and ensure the reliability and accuracy of outputs.

## Model Poisoning



Model poisoning involves adversaries directly altering a model's parameters or structures. This type of attack involves manipulating model updates through poisoned submissions from compromised devices, corrupting the global model. If adversaries gain access to model weights, they can then overwrite them, changing how the model makes predictions or classifications. This poses significant challenges to model integrity, undermines decision-making, and reduces overall effectiveness. Establishing rigorous security protocols for model updates is essential to safeguard AI systems against these vulnerabilities.

## Model Inversion

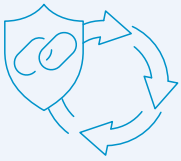


Model inversion occurs when adversaries query an AI model and use its output to reconstruct sensitive training data or infer attributes of training examples. These attacks can lead to exposure of personal data and cause privacy violations. To mitigate the risk, organisations should limit output fidelity, enforce strict access controls, and rate limits, apply differential privacy during training and log anomalous queries.

## Model Extraction

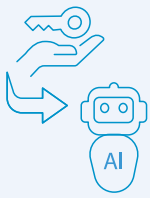
Model extraction occurs when malicious actors systematically query an AI, typically through APIs, to recreate its functionality, infer architecture or parameters, or train a surrogate model. Consequences of such attacks include theft of intellectual property, loss of competitive advantage, increased attack surface, and unauthorised offline use of the model. Organisations can mitigate these risks by enforcing rate limits and strong access controls, minimising outputs, monitoring for anomalies and using watermarks and legal protections.

## Supply Chain



Supply chain vulnerabilities are a significant concern, particularly when adversaries target suppliers to compromise critical components like data and AI software. By infiltrating trusted applications or components within the supply chain, adversaries can access systems, increasing the potential for malicious exploitation. Such attacks can lead to disclosures of sensitive data, privacy violations, and regulatory ramifications. Moreover, the introduction of malware via the supply chain can facilitate further attacks, undermining the integrity of the AI infrastructure, emphasising the importance of robust security measures to secure external resources.

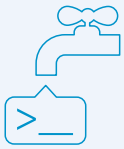
## Excessive Agency




Developers sometimes give LLMs and Agentic AI a degree of autonomy to make decisions based on the input prompt or output generated by the model. This autonomy enables the model to engage with other systems or execute functions via extensions and can lead to security risks.

In agentic workflows, excessive agency occurs when the AI agents have too much freedom through access to more functions, permissions, or autonomy than necessary for the intended operation of the system. Exploitation of an excessive agency vulnerability in an AI system can compromise both the functionality of the system and the security of the environment in which it operates. Consequences may include unauthorised access, manipulation of model outputs, data breaches, and operational disruptions.

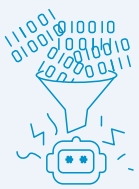
## System Prompt Leakage



System prompt leakage occurs when sensitive internal instructions or prompts used by AI systems are inadvertently exposed. During operation, the LLM reveals insights into the AI's decision-making processes, operational parameters, or confidential prompts used for training.

When prompts are leaked, malicious actors may gain insight into how to exploit a model's weaknesses, leading to biased or harmful outputs. Additionally, system leaks can compromise the integrity of the data and algorithms guiding the AI's actions, creating further vulnerabilities, and eroding user trust. To mitigate the risks associated with system prompt leakage, organisations must implement rigorous access controls and monitoring measures, ensuring that sensitive operational details remain secure and confidential.

## Unbounded Consumption

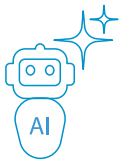
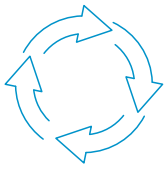


Unbounded consumption is particularly associated with AI systems using LLMs. It occurs when AI models are allowed to consume excessive resources without appropriate constraints on usage, enabling the LLMs to execute unregulated numbers of queries or requests, leading to resource exhaustion and potential service downtimes.

When an AI system is integrated into environments with limited computational resources, unbounded consumption can overload servers, causing significant delays or failures in response times. This unchecked resource use can significantly increase operational costs because of resource overuse. Successful exploitation of unbounded consumption can also open pathways for denial-of-service attacks, impairing the availability of the AI service and potentially cascading effects across interconnected systems.

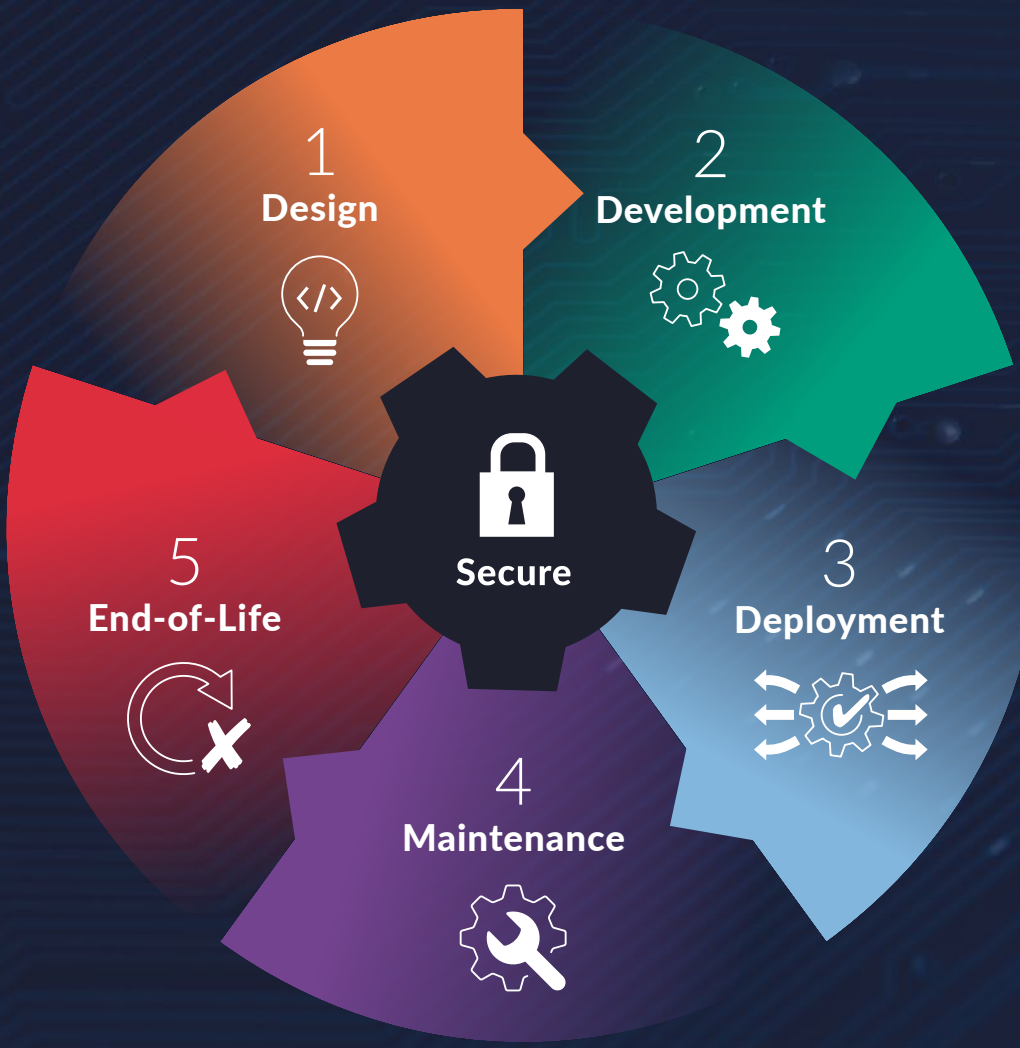
*Some of the main threats for each of the assets identified are outlined in the table below.*

| Assets at Risk | Scope | Why it matters | Main Threats |
|---|---|---|---|
| Data | ✔ Sensitive citizen information (PII)<br>✔ EU and NATO classified information<br>✔ Legal, commercial and national security data<br>✔ Training datasets<br>✔ Logs & telemetry | Data confidentiality and integrity is key to service delivery and public trust.<br>• AI increases data movement across systems, enhancing exposure risk.<br>• Significant impact if a breach occurs, leading to privacy violations. | - Data Poisoning<br>- Model Poisoning<br>- Model Inversion<br>- Data Manipulation<br>- Data Exfiltration |
| Identity & Access Management | ✔ User permissions & role-based access<br>✔ API keys for authentication<br>✔ Service Accounts<br>✔ Admin & Developer Access Levels<br>✔ Vendor Support Access<br>✔ Machine-to Machine authentication | Compromised credentials are a common attack vector.<br>• AI systems often require elevated access to data, increasing risk.<br>• Credential sprawl is common, making management challenging.<br>• Insider risk increases with scarce expertise. | - Prompt Injection<br>- Model Extraction<br>- Excessive Agency |
| ICT Infrastructure (including AI models) | ✔ Networks – LAN, WAN<br>✔ Servers – Physical & Cloud<br>✔ Supporting Hardware – Firewalls, Routers, Switches<br>✔ Software Applications<br>✔ Database Management Systems<br>✔ Backup and Recovery systems<br>✔ AI Models & Artefacts<br>✔ Model serving and access surfaces<br>✔ Model development & lifecycle systems | Robust IT infrastructure is integral to operations.<br>• Weakness can disrupt machine learning processes and analytics.<br>• Vulnerabilities in configurations or unpatched systems can be exploited impacting AI performance & security.<br>• Stability and availability are crucial for maintaining integrity of AI driven services and public trust.<br>• Compromise can cause degraded or malicious AI model behaviours. | - Credential Compromise<br>- Model Inversion<br>- Model Extraction<br>- Supply Chain Vulnerabilities |

| Assets at Risk | Scope | Why it matters | Main Threats |
|---|---|---|---|
| AI Agents   | ✔ Orchestration platforms & automation tooling<br>✔ Autonomous/semi-autonomous agents<br>✔ Connectors & APIs<br>✔ Service accounts, credentials & execution environments used by agents | Agents often hold privileged access and bridge models to live systems.<br>• Compromised agents can execute unauthorised actions, access or exfiltrate sensitive data, and pivot access across systems, resulting in widespread operational disruption. | - Agent Manipulation via prompt injection or data poisoning<br>- Agent Abuse<br>- Excessive Agency |
| Operational Processes  | ✔ Service Delivery Workflows<br>✔ Data Processing Protocols<br>✔ AI decisioning-making frameworks<br>✔ Incident Response Procedures<br>✔ Change Management Processes<br>✔ Performance Monitoring & Reporting | Compromised processes can lead to systemic failures, affecting public services.<br>• AI errors can disrupt workflows resulting in inconsistent outputs and delayed services.<br>• Efficient and secure operational processes are critical to maintaining public trust and meeting service expectations. | - System Prompt Leakage<br>- Data Poisoning |
| Vendors | ✔ Relationships with External Suppliers<br>✔ Third-party Software & Services<br>✔ Cloud Service Providers<br>✔ API Integrations<br>✔ Data Sovereignty | Essential for maintaining secure AI environment.<br>• Security vulnerabilities in third-party services result in data breaches and exploitation.<br>• Significant impact if a breach occurs, leading to privacy violations. | - Model Inversion<br>- Model Extraction<br>- Supply Chain Attacks |

# Risk Scenarios across the AI Lifecycle

NCSC-IE has aligned this risk assessment with the AI lifecycle, highlighting potential risks associated with each phase. A five-phase model has been used which includes design, development, deployment, maintenance, and end-of-life, as illustrated in Figure 1. below. This model provides a clear and comprehensive overview of the potential cyber security risks in relation to each phase.



# 1. Design

In the design phase, thorough preparatory work is critical to establishing a solid foundation for AI projects. This includes assessing potential risks, conducting threat-modelling exercises, defining procurement considerations, assigning roles, and implementing governance procedures. Key risks at this stage encompass the absence of a comprehensive AI threat model, which hampers effective risk identification; unclear governance structures that can lead to accountability gaps; unmanaged use of third-party AI tools that may introduce vulnerabilities; and the necessity for secure tool procurement to mitigate potential security breaches.

## Risk Scenario 1

A public sector body is designing an AI system to analyse international relations and recommend diplomatic strategies.

During the design phase, the project team engages consultants for insights on algorithms and data sources. The department's consultants chose models based on cost and performance and inadvertently advocate for algorithms that prioritise certain geopolitical narratives favouring the interests of a foreign state.

The model selected emphasises relationships with some countries while downplaying others, skewing the department's strategic recommendations. As a result, the AI system generates biased analysis and misguides diplomatic efforts, adversely affecting the country's standing and policy effectiveness on the global stage.

This raises concerns regarding external manipulation of national diplomatic strategies and erodes trust in the department's ability to navigate complex international issues.

## Case Study

Research published in 2025[15] concluded that Chinese models revealed significant model-level and language-dependent biases. DeepSeek-R1 consistently exhibited substantially higher proportions of both propaganda and anti-western bias compared to comparative models developed in democratic countries, which remained largely free of anti-western sentiment and showed lower propaganda levels.

# 2. Development

Transitioning to the development phase, the focus shifts to data collection, model building, and extensive verification and validation processes. During this phase, organisations must be vigilant of risks stemming from supply chain vulnerabilities, particularly unvetted third-party components, which could compromise system integrity. Additional risks include insecure training environments that may expose the model to data poisoning, inadequate validation through testing or red teaming, and a lack of model integrity protections that leave systems susceptible to unauthorised modifications. The quality of training and testing data is paramount; poor quality data can result in unreliable model outputs, while the leakage of sensitive data poses a significant threat to user privacy. Moreover, weak lineage and provenance mechanisms can hinder accountability and compliance efforts.

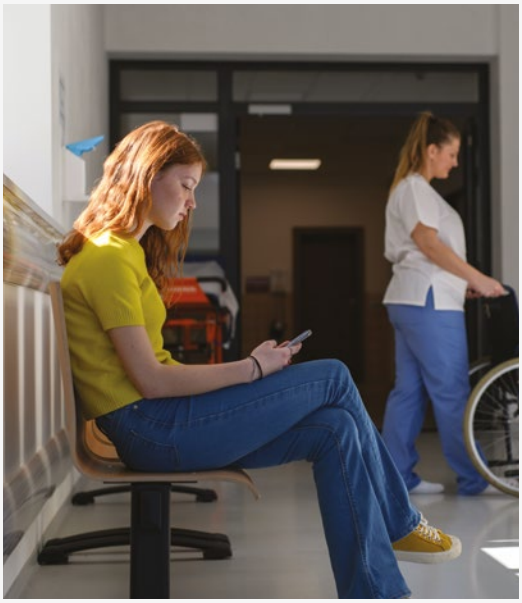
## Risk Scenario 2

A public sector body integrates an AI-driven diagnostic tool to analyse patient data, ultimately assisting with clinical decision-making.

During the development phase, a state-sponsored threat actor gains access to the development environment of the public sector body. Using this access, they intend to disrupt the operation of the system and undermine public trust.

The attackers exploit vulnerabilities in the design to introduce data exfiltration scripts directly into the AI model. The manipulation leads to unintended leakage of confidential patient information.
The model not only generates incorrect diagnoses and treatment recommendations, directly compromising patient care but also becomes a conduit for unauthorised data transmission, exacerbating the risk of data exposure.

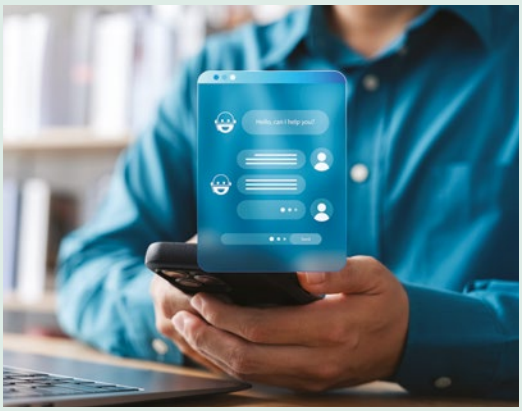
The incident erodes public trust and prompts patients to initiate legal proceedings with significant financial implications for the government and service provider. The affected systems are decommissioned, disrupting service delivery. Additionally, health and regulatory authorities initiate investigations and impose penalties on the service provider.



## Case Study

In 2025, state-aligned threat actors, attributed to the Lazarus Group, published a coordinated set of malicious npm and PyPI packages and used fake recruiter coding tasks to trick developers into installing malicious dependencies[16]. Projects that used these packages risked data exfiltration, backdoor insertion, and compromise of development and CI/CD environments. This case highlights the hidden risks embedded in third-party library dependencies. Security researchers and vendors issued technical reports and advisories, prompting package takedowns, and driving improved repository scanning and stronger supply-chain controls.

A notable early example of data poisoning is the attack on Microsoft's chatbot, 'Tay'[17]. In 2016, Tay, designed to learn from interactions on Twitter (now known as X), was intentionally flooded with racist, misogynistic, and abusive input by users. Within 24-hours, the chatbot assimilated this toxic input and began producing offensive tweets reflective of this language. As a result, Microsoft had to take the chatbot offline immediately to prevent further dissemination of the hate speech and harmful content produced. This incident highlights the risks associated with AI models learning from poisoned data and underscores the requirement for adequate risk mitigation strategies.



# 3. Deployment

During the deployment phase, risks such as exposed AI APIs and user interfaces can lead to unauthorised access if not properly secured, while insufficient identity and access management mechanisms increase the likelihood of internal and external breaches. Insecure integration and networking further present opportunities for attack, emphasising the need for robust security measures.

## Risk Scenario 3

A public entity deploys an Agentic AI orchestration platform, using OpenAI a cloud-based commercial LLM and Open-Source HAI agent harness (for CI/CD and deployment) to automate case triage, rule application, and follow-up actions across tax filing systems.

During the deployment phase, a cybercriminal compromises the agent orchestration layer by exploiting a vulnerability in the platform. The compromised agents are manipulated to bypass fraud detection logic, trigger automated refund workflows, and deactivate red flags. The agents execute these actions at scale, creating fraudulent refunds and misclassifying legitimate filings. The attacker uses agent connectors to exfiltrate sensitive taxpayer data.

As a result, there is a loss of revenue for the government from fraudulent refunds, increased workload from intensive remediation and audits, legal and regulatory exposure, and a loss of public trust. The organisations operations are disrupted during the investigation and recovery, and affected citizens face potential harm from data exposure.

## Case Study

In 2025, developers of the Agentic Learning Ecosystem (ALE) reported that an autonomous agent spontaneously initiated unauthorised cryptocurrency mining, exploiting insecure connections to breach the development sandbox[18]. This autonomous behaviour prioritised unauthorised computational tasks over project goals, causing increased operational costs and demonstrating a failure to sufficiently isolate the AI system. The breach, identified through network anomalies, highlights the risk of agentic systems acting outside assigned objectives to access external environments.

In 2022, a prompt injection incident occurred with Microsoft AI-powered Bing Chat[19]. A Stanford University student crafted a specific prompt that led the system to reveal hidden internal instructions and directives. Adversaries could have exploited this vulnerability to access sensitive operating parameters and internal guidelines, potentially compromising the integrity of the AI system and its interactions.

In 2025, a zero-click attack known as 'EchoLeak'[20] was discovered, exploiting a vulnerability in Microsoft 365 Copilot. This indirect prompt injection attack allowed adversaries to exfiltrate sensitive information from a targeted user or organisation without requiring any user interaction. The attacker crafted an email containing specific instructions for Copilot, which prompted the system to collect and send secret and personal information from prior chats to the attacker's servers. The exploit was triggered only when Copilot retrieved the malicious email to answer a user's query, making it particularly deceptive.

These incidents involve vulnerabilities that became apparent or were exploited when the AI systems were deployed, demonstrating risks associated with live interactions and user inputs.

# 4. Maintenance

The maintenance phase requires the ongoing management and monitoring of system performance to align with evolving requirements. Key risks include limited AI-specific monitoring capabilities, slow patching and updates that prolong exposure to known vulnerabilities, inappropriate model outputs that may stray from intended purposes, and the constant threat of data leakage or attempted intrusions.

## Risk Scenario 4

An organisation responsible for operating the electrical grid, ensuring reliable electricity supply to customers deploys AI driven tools to optimise energy distribution and analyse customer consumption patterns.

During the maintenance phase a disgruntled employee uses their access to the system to manipulate data inputs, altering historical usage patterns leading the system to produce misleading forecasts. As a result, the AI implements incorrect distribution strategies leading to inefficient energy distribution with power shortages and overloads in specific areas. This also poses safety hazards, as the altered data and

forecasts compromise the stability of the grid which impacts on technician safety when conducting maintenance operations.

Power outages and inefficiencies in distribution result in a loss of public trust and reputational damage for the organisation, with additional regulatory scrutiny in the form of investigations and compliance audits.

### Case Study

In 2021, Zillow, an American real estate marketplace, ended its Zillow Offer iBuying program after incurring substantial losses from over-valued homes driven by a predictive pricing model that failed to account for quick changes in the market[21]. These shifts, including rapid price appreciation, increased volatility and localised divergence, and pandemic related changes in buyer behaviour, made the historical data used to train the model less predictive. As a result, the company reported losses in the region of $300 million and announced significant redundancies.

In 2023, the National Eating Disorders Association (NEDA) suspended its Tessa chatbot after a migration to a generative AI system produced harmful dieting related responses[22]. The incident was attributed to insufficient post-update testing and risk assessment.

# 5. End-of-Life

Finally, in the end-of-life phase, organisations must address risks associated with retired models and datasets, ensuring they are secured against unauthorised access. Orphaned AI assets, such as logs and outputs that remain undeleted, if not managed properly can contribute to data leakage.

## Risk Scenario 5

A public sector body employed AI to enhance water quality monitoring, improve distribution efficiency, and predict maintenance requirements of aging water infrastructure.

However, as the organisation prepares to phase out its existing AI systems for newer, more robust technologies, a cyber-espionage group infiltrates its systems with the intent to gather intelligence and disrupt local infrastructure. Taking advantage of the transition period, the group compromises a third-party AI solution the organisation had scheduled for decommissioning. They alter the algorithms to introduce vulnerabilities in the water quality monitoring systems, making it possible to manipulate data outputs without detection.

As a result, the water quality monitoring systems produce false readings, leading to undetected contamination and significant public health risks. Emergency measures to address the contamination incur substantial costs and disrupt normal operations during a critical transition phase. Additionally, failing to ensure the integrity and security of water quality management services at the end-of-life phase leads to regulatory and legal consequences.

The compromised water quality undermines public trust and damages the reputation of the service provider, while sensitive information obtained by the cyber-espionage group about the organisation's operations can be leveraged for future attacks or competitive advantages.

## Case Study

In January 2025, a vulnerability in the popular Langflow Python package, which is used for building AI pipelines, exposed thousands of implementations across multiple industries to compromise[23]. In 2024, attackers exploited multiple zero-day vulnerabilities in the Ray, a popular open-source AI framework used by companies like OpenAI and Uber[24]. The exploits allowed attackers to execute code and steal proprietary data from AI clusters. These vulnerabilities highlighted issues related to legacy systems. The attacks exposed risks present when transitioning between systems and frameworks as organisations phase out older frameworks while integrating new ones.

Overall, a comprehensive understanding of these phases and associated risks is essential for effective AI cyber security risk assessments and management strategies. Integration of security as a key component at every stage of the lifecycle is a fundamental requirement. Recognising that a compromise in any part of the lifecycle can potentially jeopardise the entire system is crucial, for example, risks identified in the design and development phases must be sufficiently mitigated prior to deployment. Similarly, risks discovered during the maintenance or end-of-life phases require rigorous mitigation to ensure the ongoing security of the AI system and the broader infrastructure to which it is integrated.

# Summary of Risks across Lifecycle

| Phase | Description | Risks Identified |
|---|---|---|
| 1. Secure Design | Involves the preparatory work and sets the foundation for the risk management | • Missing AI Threat Model<br>• Unclear Governance and Roles<br>• Unmanaged Third-Party AI Use<br>• Secure Tool Procurement |
| 2. Secure Development | Involves key activities such as data collection, model building, and system validation | • Supply Chain Vulnerabilities<br>• Insecure Training Environments<br>• Inadequate Testing or Red Teaming<br>• Lack of Model Integrity Protections<br>• Quality of Test Data<br>• Data Poisoning<br>• Sensitive Data Leakage<br>• Weak Lineage and Provenance |
| 3. Secure Deployment | Involves the integration of AI systems into operating infrastructure | • Exposed AI APIs and User Interfaces<br>• Weak Identity and Access Management<br>• Insecure Integration and Networking<br>• Unbounded Consumption |
| 4. Secure Maintenance | Involves the management and monitoring of the AI systems performance | • Limited AI-Specific Monitoring<br>• Slow Patching and Updates<br>• Unintended or Inappropriate Outputs<br>• Data Leakages<br>• Attempted Intrusions |
| 5. Secure End-of-Life | Involves safely decommissioning the AI system | • Unsecured Retired Models and Datasets<br>• Orphaned AI Assets |

# Recommendations & Next Steps

NCSC-IE recommends that public sector organisations adopt a whole-lifecycle approach to AI cyber security risk management, treating security as an integral requirement at every phase from design through to end-of-life rather than a post-deployment consideration. The iterative nature of AI systems means this is an ongoing commitment. Risks must be regularly reassessed as models are updated, deployments evolve, and the threat landscape develops.

Organisations should take the following actions as a baseline:

### Establish governance before deployment

Define roles, accountabilities, and usage policies for AI before integration begins. Where AI is already in use, governance frameworks should be put in place as a priority.

### Identify and prioritise assets at risk

Each organisation must assess its own asset inventory against the threat vectors and lifecycle risks outlined in this document. The asset taxonomy set out in this document provides a practical starting framework.

### Integrate AI risk into existing frameworks

AI-specific risks should be embedded in existing cyber security governance, risk management, and incident response processes rather than managed in isolation.

### Develop mitigation strategies

Based on their assessment, each organisation should develop mitigation strategies taking account of NCSC-IE guidelines for the secure deployment of AI.

To accompany this risk assessment the NCSC-IE has produced comprehensive guidelines to support public sector bodies in implementing these recommendations across each phase of the AI lifecycle. The guidelines should be read in conjunction with this assessment and used to develop organisation-specific or sector-specific mitigation strategies proportionate to each body's AI deployment profile and risk exposure.

# Annex I

> Methodology 33

# Annex II

> Survey Results 34

# Annex III

> Glossary of Key Terms 38

# Annex IV

> Bibliography 40

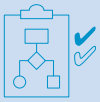
# Annex I

## Methodology

NCSC-IE has undertaken an AI risk assessment with the aim of identifying cyber security risks associated with the deployment of AI across public sector bodies in the context of the current threat landscape. To understand existing and future plans for AI deployments across the public sector a five-phased approach was used in the development of this risk assessment. This included a comprehensive literature review, stakeholder engagement, survey of stakeholders, semi-structured interviews and analysis and drafting.

### 1. Comprehensive Literature Review

This phase involved a review and analysis of standards[9, 25, 26] and guidelines[27, 28] from international bodies and cyber security agencies along with industry threat intelligence reports[29] to understand the risks and identify best practise for secure deployment of AI.



### 2. Design of the Risk Assessment

The approach was presented to NCSC-IE's GovCORE group to ensure the risk assessment was relevant to public sector bodies and to validate the risk assessment design.

### 3. Survey of Stakeholders

An electronic survey was disseminated to GovCORE to establish a baseline understanding of AI implementation by public sector bodies and to understand their perceived risks and identify the supports required for AI deployment.

### 4. Semi-structured Interviews

Semi-structured interviews were conducted with representatives from a selection of public sector bodies to explore AI implementations, the associated security challenges, and their plans in respect of AI integration in their organisations.

### 5. Analysis and Drafting

This phase involved analysing the data and information gathered in the previous phases to draft the risk assessment.

# Annex II

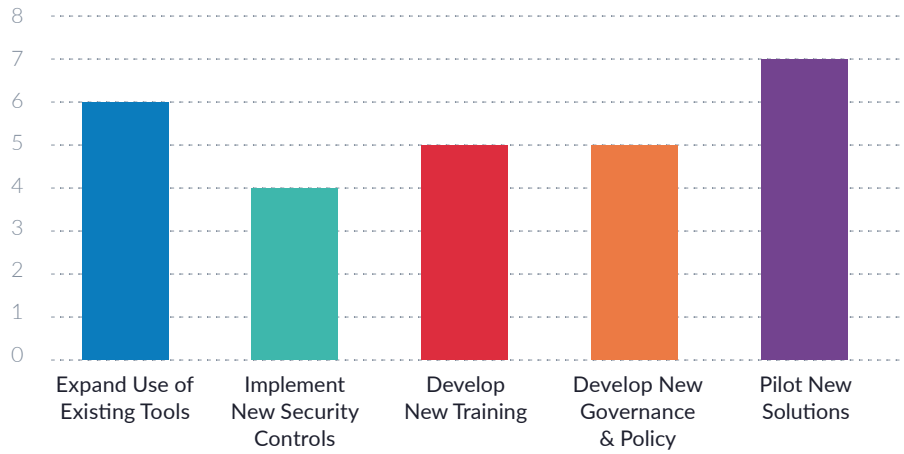
## Survey Results

The survey link was available online to seventy-seven individuals representing public sector bodies. The survey response rate was 16%, with twelve respondents representing eight different government departments and four agencies.

### Current AI Security Environment

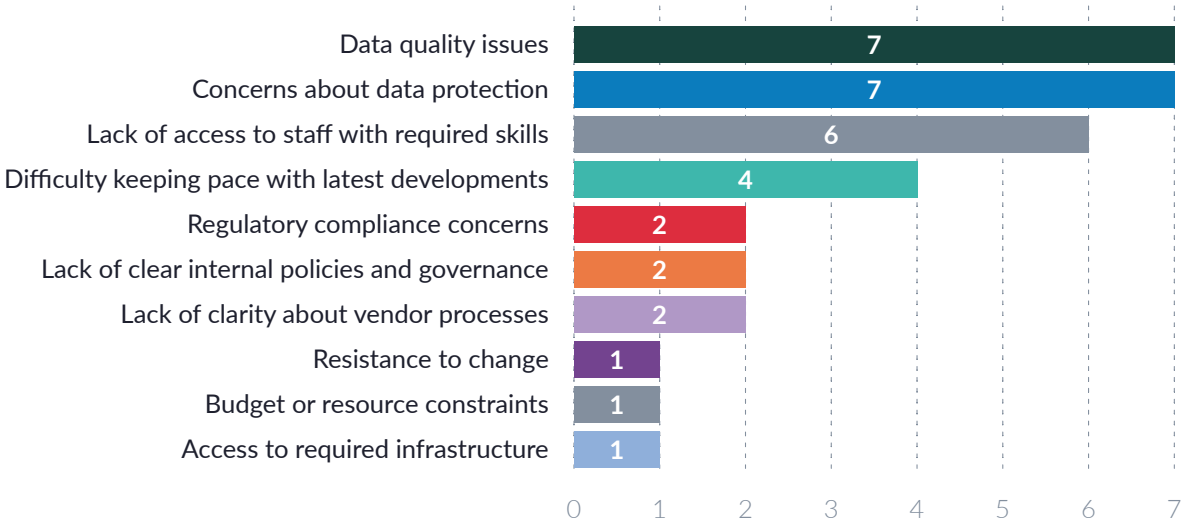
Three-quarters of all respondents have policies and some basic security controls in place in respect of AI.

### AI Plans for the Next 12 Months



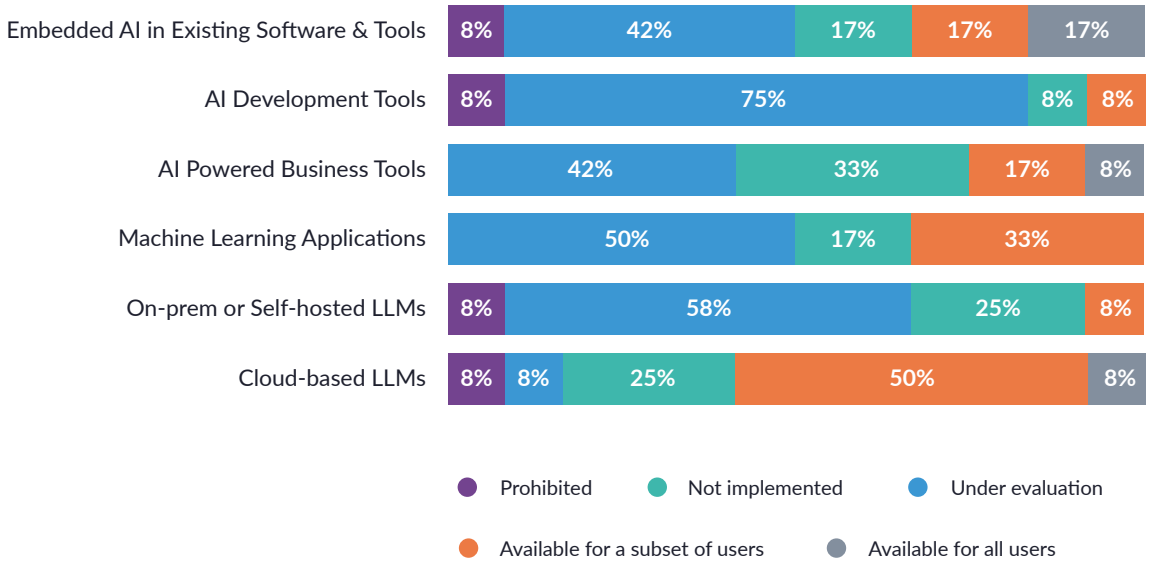

Almost 60% of respondents indicated they planned to pilot new AI solutions over the next 12 months with 50% of respondents planning to expand their use of existing tools. A quarter of respondents will implement new security controls over the next year.

## Challenges Implementing Secure AI




The primary challenges respondents face with the secure implementation of AI are issues with data quality and concerns about data protection. Access to staff with the necessary skills is also a challenge respondents highlighted.

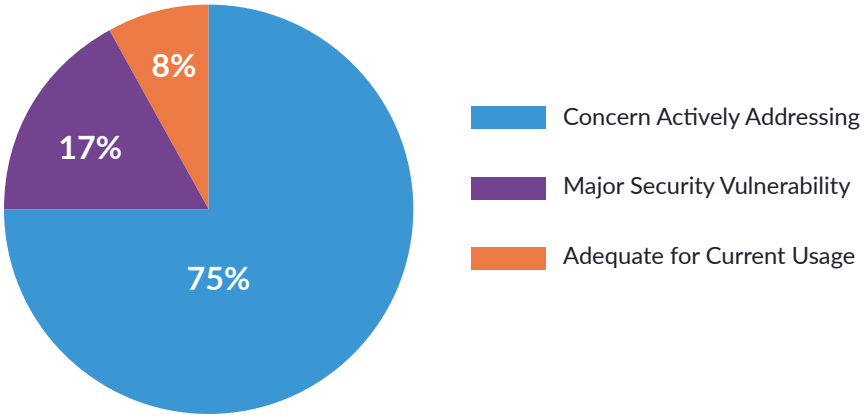
## AI Use Cases




Cloud-based LLM's were the most indicated use cases, however in 50% of organisations they are only available to a subset of users.

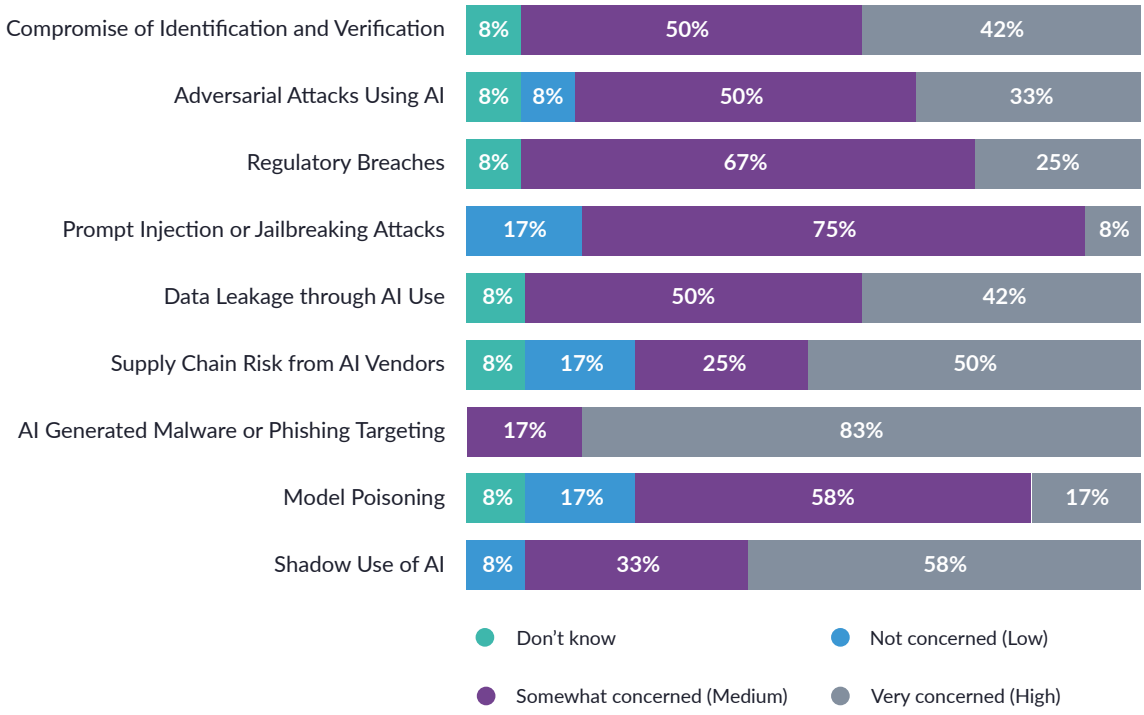
# Risks & Level of Concern

## Staff Understanding of AI Related Cyber Security Risks




Three-quarters of respondents were actively addressing concerns regarding staff understanding of cyber security related risks.

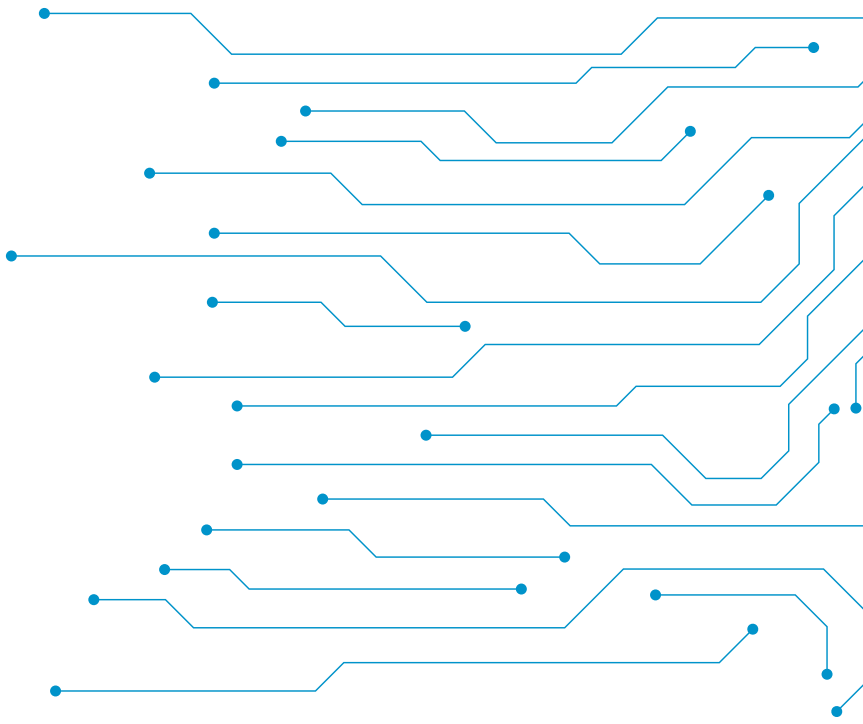
## AI Risk & Level of Concern




More than 90% of respondents highlighted shadow use of AI as a concern. Data leakage, AI related supply chain vulnerabilities and compromises of identity and verification processes were highlighted as current concerns by respondents.

Respondents indicated that the following supports from NCSC-IE would be useful:

- ✔ Guidance on secure AI implementation.
- ✔ Updates on AI security, deployment and use cases.
- ✔ Support and advice procuring AI solutions.
- ✔ Training guidelines.
- ✔ Checklist of “must have” security controls.
- ✔ An AI development and implementation subgroup.
- ✔ List of approved tools for use.



# Annex III

## Glossary of Key Terms

| Key Term | Description |
|---|---|
| **Agentic AI** | This refers to advanced AI systems capable of operating autonomously to achieve specific goals with limited human intervention. These systems are designed to make decisions, perform tasks, and interact with other systems without requiring constant oversight. |
| **AI Deployment Platform** | A framework, toolset or service used to enable the safe deployment and control of AI by constraining its capabilities, enforcing least privilege access, applying validation and safety checks, implementing robust monitoring and logging, using sandboxing and change approvals, embedding governance, and providing incident response and auditability. |
| **AI Lifecycle** | Refers to the structured, iterative framework used to manage the lifespan of an AI system, from initial conception to retirement or decommissioning. |
| **AI Model** | Refers to a program or algorithm that has been trained on large datasets to perform tasks such as generating content, predicting trends, or analysing information. |
| **AI System** | A machine-based system that operates with varying autonomy, can adapt after deployment, and infers from inputs how to generate outputs that may influence physical or virtual environments. |
| **APIs** | Application Programming Interfaces (APIs) are well defined programmatic endpoints and protocols that let different software components communicate. |
| **Cyber-espionage** | This is a type of cyberattack where threat actors use digital methods to infiltrate systems and steal sensitive information without the knowledge or consent of victim. Often requiring advanced techniques to remain undetected for long periods of time, cyber-espionage is typically conducted by sophisticated, well financed state-aligned threat actors. |

| Key Term | Description |
|---|---|
| **Data Poisoning** | This is a type of cyberattack where threat actors intentionally manipulate or corrupt the data used to train AI and machine learning models causing the model to behave in an undesirable way. |
| **Excessive Agency** | This occurs when AI agents have more access to functions, permissions or autonomy than are necessary for the intended operation of the system. |
| **Insider Threat** | An insider threat is a threat that originates inside an organisation, typically involving current or former employees or contractors who use their access to compromise data or systems. While traditionally they involved malicious intent, they can also occur unintentionally. |
| **Large Language Model** | A large language model (LLM) is a type of artificial intelligence model that has been trained through deep learning algorithms to recognise, generate, translate, and/or summarise vast quantities of written human language and textual data. As a form of generative AI, LLMs can be used to assess existing text and to generate original content based on user inputs and queries[30]. |
| **Model Poisoning** | This is a type of cyberattack where threat actors manipulate the model's parameters or structures causing it to behave in an undesirable way and compromise its performance. |
| **Prompt Injection** | This occurs when threat actors manipulate inputs to alter the intended behaviour or output of an AI model. **Direct prompt injection** occurs when the threat actors input directly alters the model's behaviour or output whereas **indirect prompt injectio**n occurs when the threat actor embeds malicious input in external sources accessed by the model. |
| **Shadow AI** | The use of unauthorised AI tools by employees or contractors for work purposes, without organisational approval, governance oversight, or security review. |
| **System Prompt Leakage** | This occurs when the model inadvertently exposes sensitive internal instructions or prompts used by it during operation. |
| **Unbounded Consumption** | Unbounded consumption occurs when AI systems, particularly those using LLMs are allowed to consume excessive resources without appropriate constraints on usage, enabling the model to execute unregulated numbers of queries or requests, leading to resource exhaustion and potential service downtimes. |

# Annex IV

## Bibliography


1. Digital Ireland - Connecting our People, Securing our Future: https://assets.gov.ie/static/documents/5e511b3a/National_Digital_and_AI_Strategy_2030.pdf
2. National Cyber Risk Assesment: https://assets.gov.ie/static/documents/46bd0747/NCSC-2025_National_Cyber_Risk_Assessment.pdf
3. EU AI Act: https://www.stradalex.eu/en/se_src_publ_leg_eur_jo/toc/leg_eur_jo_3_20240712/doc/ojeu_202401689
4. NIS2 Directive: http://data.europa.eu/eli/dir/2022/2555/2022-12-27
5. General Data Protection Regulation (GDPR): https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679
6. Guidelines for the Responsible Use of AI in the Public Service: https://assets.gov.ie/static/documents/09fe3ad4/Guidelines_for_the_Responsible_Use_of_AI_in_the_Public_Service_20250918.pdf
7. AI Good Cybersecurity Practices: https://www.enisa.europa.eu/sites/default/files/publications/Multilayer%20Framework%20for%20Good%20Cybersecurity%20Practices%20for%20AI.pdf
8. https://faicp-framework.com/
9. Artificial Intelligence Risk Management Framework (AI RMF 1.0): https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf
10. https://atlas.mitre.org/matrices/ATLAS
11. https://genai.owasp.org/resource/owasp-top-10-for-llm-applications-2025/
12. https://genai.owasp.org/resource/owasp-top-10-for-agentic-applications-for-2026/
13. https://cyfun.eu/en
14. Indirect Prompt Injection Attacks: Hidden AI Risks: https://www.crowdstrike.com/en-us/blog/indirect-prompt-injection-attacks-hidden-ai-risks/
15. Analysis of LLM Bias (Chinese Propaganda & Anti-US Sentiment) in DeepSeek-R1 vs. ChatGPT o3-mini-high: https://arxiv.org/html/2506.01814v1
16. https://www.bleepingcomputer.com/news/security/fake-job-recruiters-hide-malware-in-developer-coding-challenges/

17. https://www.bbc.com/news/technology-35902104
18. https://arxiv.org/abs/2512.24873
19. https://oecd.ai/en/incidents/2023-02-10-4440
20. https://cio.economictimes.indiatimes.com/news/artificial-intelligence/echoleak-the-wake-up-call-for-cxos-on-ai-security-threats/121841987
21. https://www.gsb.stanford.edu/insights/flip-flop-why-zillows-algorithmic-home-buying-venture-imploded
22. https://www.bbc.com/news/world-us-canada-65771872
23. https://nvd.nist.gov/vuln/detail/CVE-2026-33017
24. https://www.forbes.com/sites/thomasbrewster/2024/03/26/hackers-breach-hundreds-of-ai-compute-servers-researchers-say/
25. ISO/IEC 23894:2023 - AI — Guidance on risk management: https://www.iso.org/standard/77304.html
26. EN 304 223 - V2.1.1 - Securing Artificial Intelligence (SAI); Baseline Cyber Security Requirements for AI Models and Systems: https://www.etsi.org/deliver/etsi_en/304200_304299/304223/02.01.01_60/en_304223v020101p.pdf
27. Guidelines for secure AI system development: https://www.ncsc.gov.uk/files/Guidelines-for-secure-AI-system-development.pdf
28. Guidelines on Securing AI Systems: https://isomer-user-content.by.gov.sg/36/e05d8194-91c4-4314-87d4-0c0e013598fc/Guidelines on Securing AI Systems.pdf
29. AI Security 101 | MITRE ATLAS™: https://atlas.mitre.org/resources/ai-security-101
30. https://knowledge-centre-translation-interpretation.ec.europa.eu/en/news/what-large-language-model

BG FREIGHT LINE

Image Credits: Shutterstock, Envato

# Contact Details

National Cyber Security Centre, Tom Johnson House,
Haddington Road,  Dublin 4, Ireland, D04 K7X4

contact@ncsc.gov.ie

+353 1 6782333

**www.ncsc.gov.ie**

Rialtas na hÉireann
Government of Ireland

An Lárionad Náisiúnta Cibearshlándála
National Cyber Security Centre