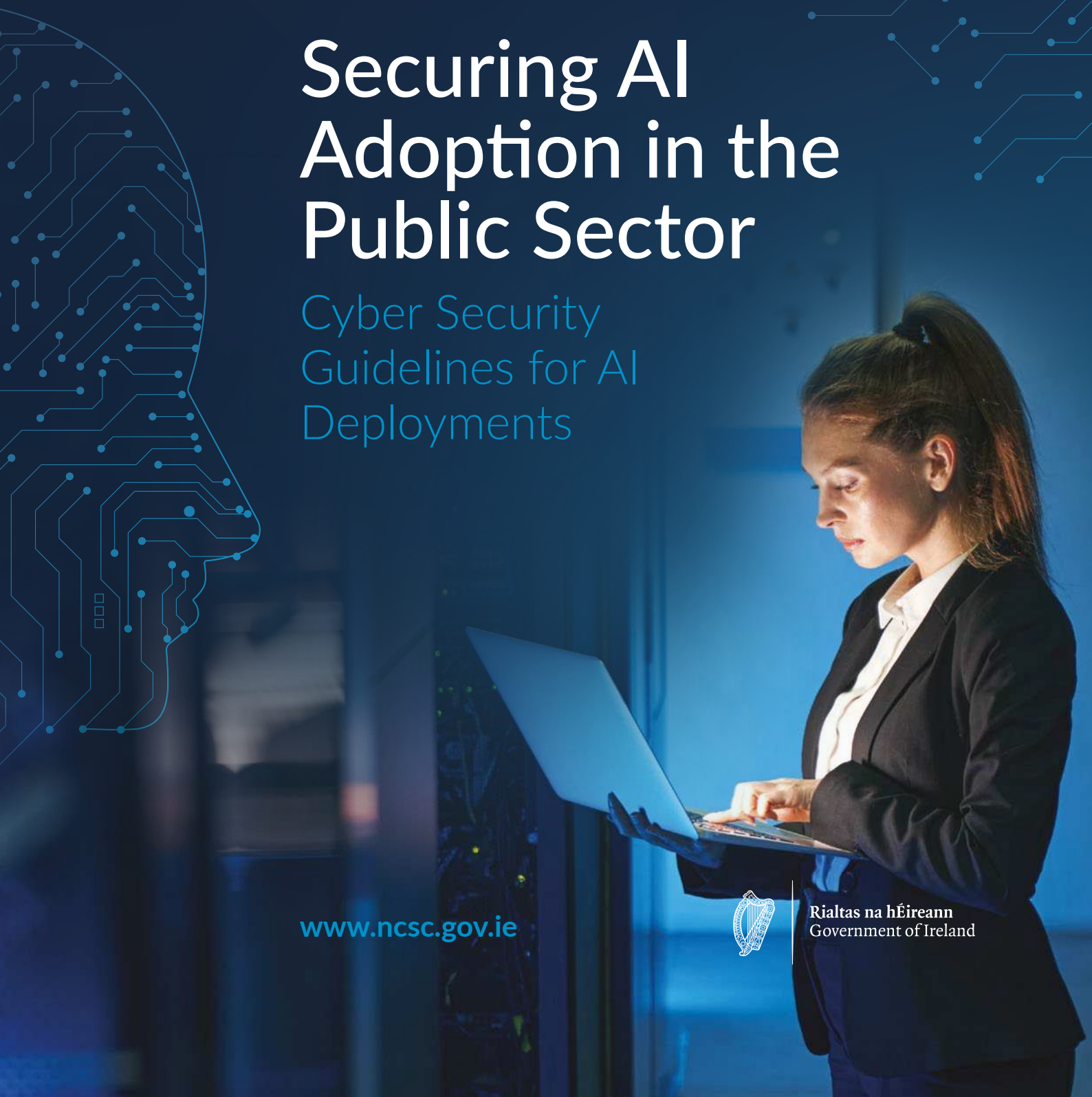
An Lárionad Náisiúnta Cibearshlándála
National Cyber Security Centre

Securing AI Adoption in the Public Sector

Cyber Security Guidelines for AI Deployments



www.ncsc.gov.ie

Rialtas na hÉireann
Government of Ireland

Executive Summary

NCSC-IE is fully supportive of AI adoption across the public sector and has developed these guidelines to help public sector bodies to prepare for the cyber security risks that come with it. The intention is to enable innovation rather than constrain it, giving public sector bodies the confidence to adopt AI on the understanding that it is being deployed securely, resiliently, and in line with regulatory obligations.

This cyber security guidance is a companion to the broader Guidelines for the Responsible Use of AI in the Public Service published by the Department of Public Expenditure, Infrastructure, Public Service Reform and Digitalisation (DPER), which set the overarching framework for how the public service should adopt and use AI. Where the DPER guidance addresses responsibility, transparency, accountability, and public good, these guidelines address the cyber security obligations that flow from that framework. They are designed to be complementary and should be used in conjunction with one another.

The NCSC-IE Guidelines are the operational companion to the NCSC AI Cyber Security Risk Assessment, which should be reviewed first, and which informed the scope and emphasis of this document throughout. It is also recommended that public sector bodies apply these principles within an overarching cyber security governance system, such as the Cyber Fundamentals Framework (CyFun), rather than as a stand-alone exercise; a mapping between the guidelines and CyFun to support their interaction is included in our associated resources.

They are aimed primarily at Chief Information Officers (CIOs), Chief Information Security Officers (CISOs), Chief Technical Officers (CTOs), and senior managers and the teams supporting who are responsible for the delivery and management of ICT services and systems across government departments, agencies, and other public sector bodies. These guidelines may also be useful for management boards who will have expanded responsibilities under NIS2. They have been tailored for Irish public sector bodies and incorporate available international standards (principally ETSI EN 304 223). It is designed to be flexible and accessible to smaller public sector bodies without specialist AI security capability and useful as a starting point for larger organisations with more mature functions.

The document is structured around seven principles spanning the five phases of the AI lifecycle: design, development, deployment, maintenance, and end-of-life.

NCSC-IE will continually assess the AI threat and capability landscape and will review its AI security guidance as the operational environment, feedback from public sector bodies, and regulatory context evolve. This is a living resource rather than a point-in-time publication.

Contents

Introduction

As AI redefines the technological landscape, public sector organisations must shift their focus from whether to adopt AI, to how to implement it securely. Ireland's Digital and AI Strategy "Digital Ireland – Connecting our People, Securing our Future" outlines a proactive approach to AI adoption supported by the DPER Guidelines for the Responsible Use of AI in the Public Service.

Digital Ireland – Connecting our People, Securing our Future

Government of Ireland

The NCSC-IE guidelines presented in this document act as a complementary resource to the NCSC Cyber Security Risk Assessment for Public Sector Deployment and have been designed to enable organisations deliver on the vision and ambition underpinning the National Digital and AI Strategy 2030.

2025 National Cyber Risk Assessment

NCSC-IE

The NCSC AI Cyber Security Risk Assessment identified a complex landscape of interconnected risks, including data security concerns, novel adversarial techniques and governance challenges. This was supported by stakeholder engagement which revealed that AI adoption across the Irish public sector is accelerating but for many organisations remains at an early adoptive stage. Security concerns remained one of the primary barriers to adoption, and public sector bodies have asked NCSC-IE for practical guidance on secure implementation.

These guidelines have been developed in response and set out seven principles linked to the AI lifecycle, and are supported by key considerations, suggested resources, and implementation scenarios drawn from realistic public sector contexts. By adhering to these principles, public sector bodies can ensure that their digital transformation is resilient, secure, and capable of meeting the evolving challenges of the modern threat landscape.

Scope

Approach:

These guidelines are advisory rather than mandatory and should be applied through a risk-based approach calibrated to each organisation's risk environment and business requirements.

Given the rapid pace of technological change in AI, this document provides a foundational level of support that will be regularly enhanced through future updates and supplementary advisory notes.

Who these guidelines are for:

These guidelines are intended for public sector organisations in Ireland that are developing, procuring, deploying, or already operating AI-enabled systems or services.

They are aimed primarily at CIOs, CISOs, CTOs, and senior managers responsible for the delivery and management of ICT services and systems across government departments, agencies, and other public sector bodies. These guidelines may also be useful for management boards who will have expanded responsibilities under NIS2. They are also applicable to other sectors who have or plan to integrate AI in their organisations.

How to use them:

They are intended as the operational companion to the NCSC AI Cyber Security Risk Assessment. The risk assessment sets out the threats facing public sector AI deployments; the guidelines set out the principles and controls that respond to those threats. Both documents should be read together. Within the guidelines, the seven principles are organised across the five AI lifecycle phases (design, development, deployment, maintenance, end-of-life), and each principle is supported by key considerations, suggested resources, and implementation scenarios.

Relationship to other guidance:

These guidelines are one element of a wider package supporting public sector AI adoption. The DPER Guidelines for the Responsible Use of AI in the Public Service set the overarching framework, including the Decision Framework for evaluating whether AI is the right solution, the Responsible AI Canvas for planning, and the AI Lifecycle Guidance Tool. The Data Protection Commission (DPC) also publishes guidance on AI and data protection covering The Law Directive obligations. The AI Office of Ireland will lead on EU AI Act implementation. CyFun remains the NCSC-IE's recommended cyber assurance framework within which these guidelines should be implemented. Together these instruments form a coherent package; public sector bodies should not read any one in isolation.

Disclaimer:

The guidelines do not represent legal advice, regulatory direction, operational approval or risk acceptance. No part of these guidelines should be interpreted as validation, endorsement, or confirmation that a particular approach, control set, or decision meets legal, regulatory, or supervisory requirements howsoever arising including but not limited to EU regulations, directives, national legislation or regulations, guidance, and/or policies.

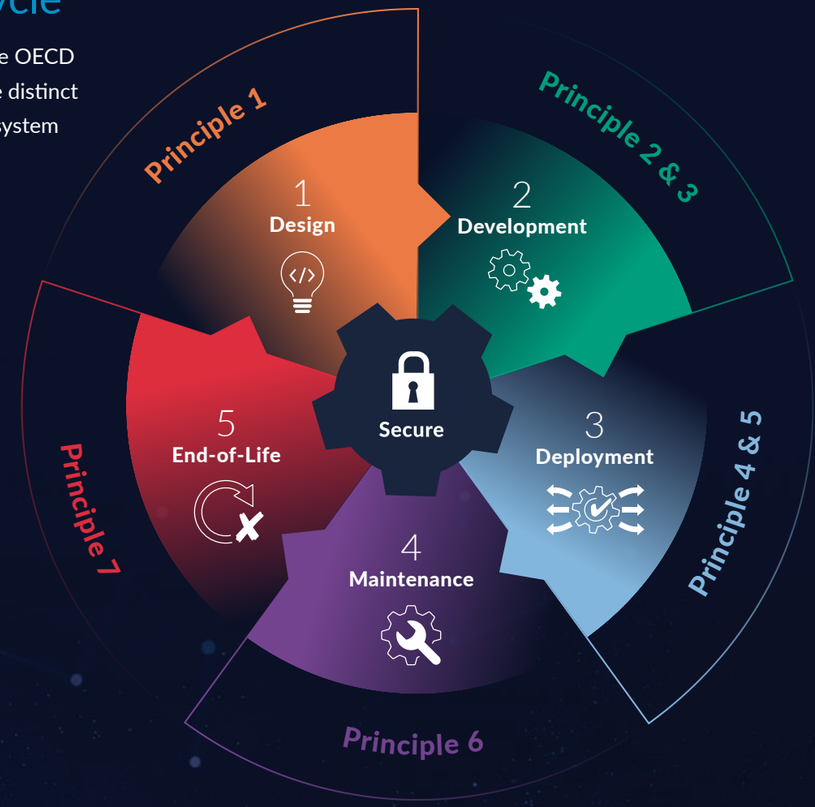
Any examples, opinions, or suggested approaches shared here are illustrative in nature and are not a substitute for your own governance, assurance processes, or independent professional advice. Organisations remain solely responsible for their own decisions (including security decisions), compliance obligations, and risk management outcomes.

AI Guidelines

The secure design, development and deployment of AI systems is a continuous process, not a one-off exercise.

AI Lifecycle

As defined by the OECD AI Principles, the distinct phases of an AI system lifecycle include:



All stages of the lifecycle should be considered, from design through to decommissioning, and it relies on active collaboration between stakeholders. Secure deployment of AI must be embedded from the beginning of the process rather than trying to reactively resolve problems created at a later stage.

How these guidelines work

These guidelines set out seven principles spanning the full AI lifecycle. Each principle is supported by a set of key considerations framed as questions. Additional resources and implementation scenarios drawn from realistic public sector contexts are also included. They are designed to be accessible to the smaller public sector body's IT and governance teams, while providing a robust starting point for larger organisations.

This document balances plain-language guidance with the technical precision required for cyber security tasks. They do not attempt to set out every control a public sector body may need, or every technical measure an advanced deployment may require. Their purpose is to give public sector bodies a logical framework they can work through, apply proportionately to their own risk environment, and build on as their AI capability matures.

Many of the actions and mitigations described are not isolated to AI. They involve applying the same controls and mitigations from existing cyber security frameworks such as the CyFun framework recommended by the NCSC-IE. The value is in applying them consistently and explicitly across the AI lifecycle and addressing AI-specific risks where necessary.

These guidelines are structured to align broadly with ETSI EN 304 223 and its associated technical specifications and requirements, as well as other relevant international guidance and best practice documentation.

Organisations seeking fuller technical implementation detail, formal conformance, or audit-ready provision-by-provision traceability should engage directly with the relevant ETSI standards, see Appendix II ETSI guidelines for further details.

AI Lifecycle

Within the AI lifecycle security must be considered and integrated throughout each stage. By adopting secure by design principles, organisations can address risks ranging from supply chain vulnerabilities to the final deletion of data at end-of-life phase.

1. Design

Security should be considered and embedded from the beginning of the design process with clear accountability, and oversight included as part of the design rather than bolted on at a later stage. This involves integrating security controls and governance measures into the initial design and architecture to ensure the system is robust and resilient. The design process is iterative, requiring continuous adjustment based on risk assessment, technical feedback and changes to the operational environment.

Principle 1: Build trust and security through informed design and risk awareness

The design phase establishes the infrastructure on which the AI system will be built. For this foundation to be secure, decision makers must possess awareness to embed essential security structures into the system from the beginning. This phase focuses on developing a proposal that clearly documents the risks alongside the benefits of the chosen approach. By maintaining effective oversight, the organisation ensures that staff are well informed and that the progression of the project is fully documented, allowing design decisions to align with the organisation risk profile.

Key Considerations

Training and awareness

- Has the organisation established tailored AI security awareness programmes for staff involved in the design, development, procurement, and/or operation of AI systems, including non-technical staff with decision-making responsibility?
- Is there a requirement for all staff using AI tools to complete AI literacy training before access is granted, covering acceptable use, data handling, and reporting of concerns along with the risks of Shadow AI?
- Are awareness arrangements reviewed and refreshed on a defined cadence to reflect the evolving threat landscape?

Design and risk assessment

- Is there a documented use case with a clearly identified business objective, stakeholders and a nominated accountable owner?
- Has the team completed the DPER Decision Framework and Responsible AI Canvas, and is the resulting conclusion to proceed with AI formally recorded?
- When assessing the proposal, is security as important a consideration as what the AI system will do and how it will perform?
- Has the proposal been risk assessed, considering the threat environment and risk classification of the use case and the organisation?

Human oversight

- Do your organisation's governance arrangements include formal approval before progressing to development?
- Does the cyber security architecture explicitly support and enforce the oversight levels mandated by the DPER Responsible AI framework, through embedded manual intervention points and 'human-in-the-loop' controls, where the risk profile demands it?

Data governance

- Does the project clearly state which data the system will use, and what should be out of bounds?
- Have the data owners been involved in the design decisions to ensure it can support the required outcomes?
- Has a Data Protection Impact Assessment (DPIA) been conducted in accordance with statutory requirements, with all mandatory risk mitigations integrated into the system's technical specifications prior to development?

Procurement and vendor selection

- Have the available deployment options (cloud-hosted, on-premises, and open-source or open-weight models) been evaluated against the use case, with trade-offs in cost, control, data residency, security assurance and long-term maintenance documented as part of the decision?
- Have security requirements been defined as part of the procurement specification, rather than addressed after vendor selection?
- Has the vendor's approach to data handling, data residency, model training use, and incident notification been evaluated against the organisation's risk profile?
- Does the procurement evaluation consider the vendor's own supply chain, including the underlying model provider and any sub-processor?

2. Development

In the development phase, the focus shifts to data collection, model building, testing and validation. This requires protection of assets, data provenance and supply chain integrity to build AI systems that are resistant to attack, resilient and safely integrated with existing infrastructure. While this phase primarily involves those who are developing the systems, it also creates downstream responsibilities for operators, data owners, end users and those providing oversight to ensure these protections are maintained.

Principle 2: Identify and secure all assets ensuring end-to-end traceability and recovery

Maintaining a secure environment requires full visibility and control over all system components, including AI-specific elements like training data, prompts and models. By recording the provenance and lineage of these assets, you can verify their integrity, detect tampering and quickly trace the root cause of system failures.

Key Considerations

Identification of assets

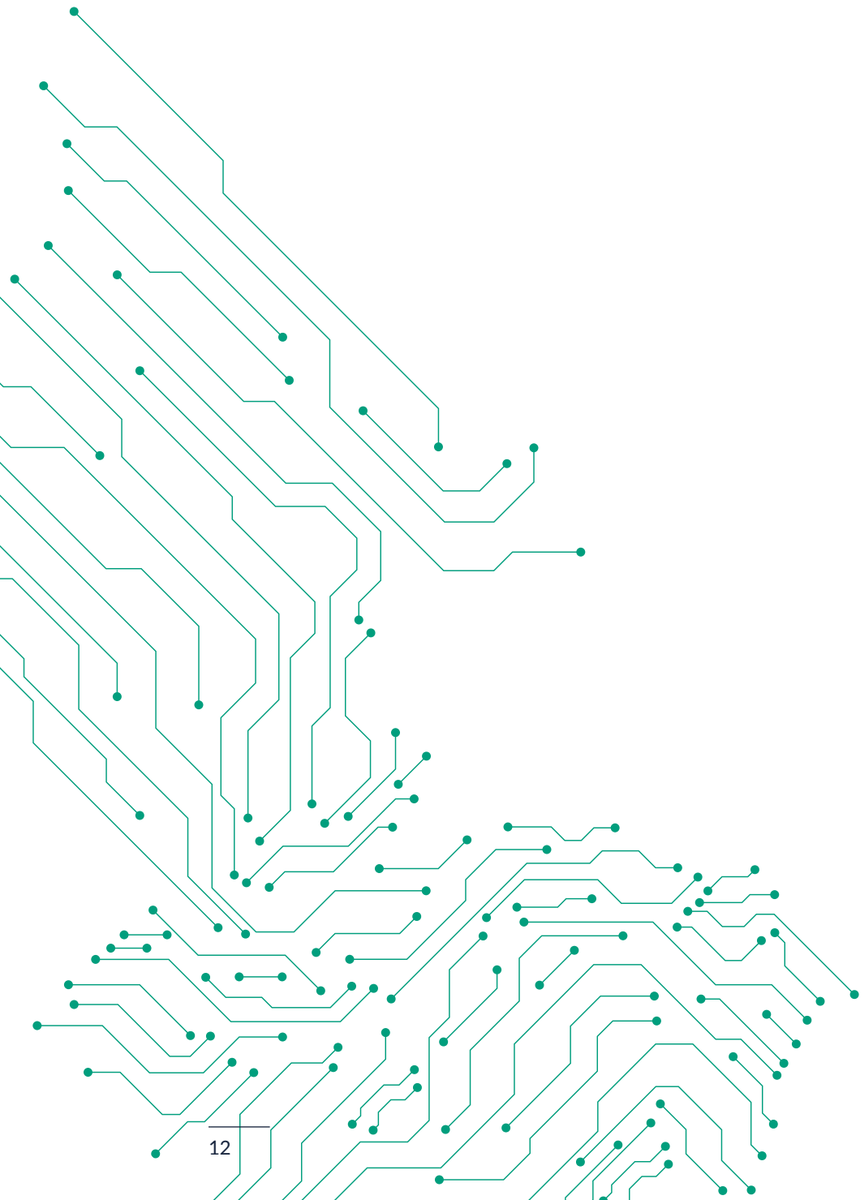
- Does your organisation maintain a documented and updated AI asset record including datasets, models, prompts, third-party library versions, connectors and transformations with provenance, version history, identified owners and a review cadence?

Traceability and auditability

- Is there a documentation package or manual for each AI system covering its purpose, data sources, training or fine-tuning methodology, prompt architecture, known limitations, and decisions taken during design?
- Can you trace the full history of every asset, including its identified owners, and any subsequent transformations or changes?
- Are there clear rules, compliant with relevant legal requirements, to ensure you only use the data you need?
- Are there clear rules for using third-party software libraries and external models, including checks of versions against an approved whitelist before they are imported into the development environment?

Integrity and recovery

- Is the AI development environment, whether cloud-based or on-premises, strictly segregated from production systems with equivalent access controls, network segmentation and logging?
- Are immutable access logs and version histories maintained for all data to allow for a rapid recovery to a known good state should an incident occur?
- Is the training data pipeline secured against injection or poisoning attacks with validation checks applied?
- Do you verify the integrity of your assets, checking all ingested assets against provenance records to detect tampering or unauthorised changes?



Principle 3: Build resilience through effective security controls and supply chain assurance

This principle focuses on implementing security controls and robust threat modelling, to protect the system from attacks. Central to this is supply chain assurance, which involves verifying the security of every external component. By integrating these requirements into the development lifecycle, the organisation ensures that AI systems are resilient to disruption and that third-party risks are mitigated.

Key Considerations

Security controls

- Has threat modelling been completed for major attack surfaces including the model, data pipeline and interfaces?
- Are encryption, key management and secrets-handling procedures defined and implemented for data, models and communications?
- Have developers received AI-specific secure development training including awareness of adversarial risks?
- Do your organisation's disaster recovery plans account for attacks on AI systems with the ability to return systems to a known good state if required?

Supply chain assurance

- Do all third-party AI agreements mandate strict security controls such as data encryption, vulnerability disclosure, incident notification timelines, and clear enforceable limits on how data will be used for model training or fine-tuning?
- Has the supplier provided verifiable evidence such as a Software Bill of Materials, (SBOM) documented patch cadence and independent assurance reports and has it been reviewed and saved securely?
- Where open-source models are used, how will the maintenance status of the upstream project be monitored, and have contingency plans to allow for rapid failover or replacement been put in place if the model is deprecated, withdrawn or compromised?

3. Deployment

This phase focuses on how to securely release AI systems, placing them into operation, providing support outside of the controlled development environments and ultimately how users engage with them.

Deployment decisions should be based on successful completion of robust testing and validation. Users should understand what the limitations of the system are and how their data (including prompt data) is retained and used.

Principle 4: Implement rigorous testing and validation to ensure reliability

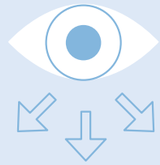
AI systems behave differently when compared to traditional software systems with predetermined logic. Their probabilistic nature can produce unexpected results. To safeguard an AI deployment, it must undergo rigorous stress testing designed to uncover hidden vulnerabilities. Clear readiness criteria must be defined before testing begins, to ensure objective decision making against predefined safety and performance standards.

Key Considerations

Testing and validation

- Have clear and documented performance benchmarks and readiness criteria been established prior to testing to guide the deployment decision?
- Has formal sign-off against these benchmarks been obtained prior to commencing deployment?
- Have rollback and fail-safe mechanisms been defined and tested to ensure the system can be safely deactivated if it behaves unexpectedly?
- Does the test plan cover functional, edge case, performance and fairness tests alongside AI-specific adversarial tests such as prompt injection, jailbreak attempts and training or context extraction where relevant to the use case?
- Has a red-team exercise or independent third-party verification been completed?
- Are training and tuning environments isolated from production systems with separate access controls and deployment pathways?
- Is there a documented pre-deployment process to verify that supplier security standards remain aligned with internal controls?
- Is there a documented process for retesting the system after any model updates or dataset changes before it goes live?

Principle 5: Support oversight and assurance through accountable, explainable and transparent operation



Accountability relies on three pillars: human oversight, explainability and transparent communication. Controls must be in place to actively supervise and, if necessary, override system actions to prevent unintended outcomes. Explainability ensures that the rationale for outputs can be interrogated to a level proportionate to the decision's impact. Transparency allows users understand what the system does, what its limits are, and how their data is handled. Robust support arrangements are essential to effectively manage the secure deployment and maintenance phases.

Key Considerations

Oversight and accountability

- Building on the accountability arrangements established under the DPER Responsible AI Framework, does the organisation have nominated trained personnel with the technical access, the authorisation, and the procedural authority to review, override, or approve system decisions during operation?
- Following the transparency and explainability commitments set under the DPER Responsible AI Framework, is the provided explainability proportionate to the system's risk classification, and does it provide sufficient technical detail for the cyber security and operational teams to understand, challenge, and respond to unexpected or malicious outputs?
- Is there adequate monitoring and alerting in place to notify operators of unusual or unexpected outputs?
- Is there a mechanism to monitor whether users are bypassing the system in favour of unauthorised AI systems and is this fed back to development?

Operational coordination

- Is there an agreed runbook with details of how your organisation and the developer will cooperate during a cyber security incident?
- Have clear roles and responsibilities, named points of contact and escalation points been documented between your organisation and the developer?
- Does the agreement include measures to contain and mitigate the impacts of a security breach?

Communication and feedback

- Have the end-user transparency commitments set out in the Responsible AI Canvas been operationalised in accessible user-facing guidance, and does the guidance accurately reflect what data the system collects, how it is used, and the security constraints that apply?
- Have end users been informed clearly about prohibited uses of the AI system in line with the policy framework set under the DPER Guidelines for the Responsible Use of AI in the Public Service, and are there technical or procedural controls in place to enforce those prohibitions where they have cyber security implications?
- Is there a clear process for end users to raise concerns when the system produces an output, they believe to be incorrect, misleading or outside its intended scope?
- Does your organisation have a coordinated vulnerability disclosure policy that covers AI systems?
- Is there a process for communicating security update notices to users when there are significant changes or vulnerabilities?

4. Maintenance

Once an AI system is deployed and live, there are ongoing maintenance obligations including continuous monitoring, patching, and comparison to previous performance to keep the system secure and flexible to changes in the threat environment.

Principle 6: Maintain secure operations through continuous oversight and rapid response

Once an AI solution is live, models can "drift" from their original performance or be targeted by adversarial techniques that weren't present during initial testing. By treating oversight incident response as an active, adaptive process, the organisation can respond rapidly to emerging threats preventing impact on the organisation or its users.

Key Considerations

Monitoring & Performance

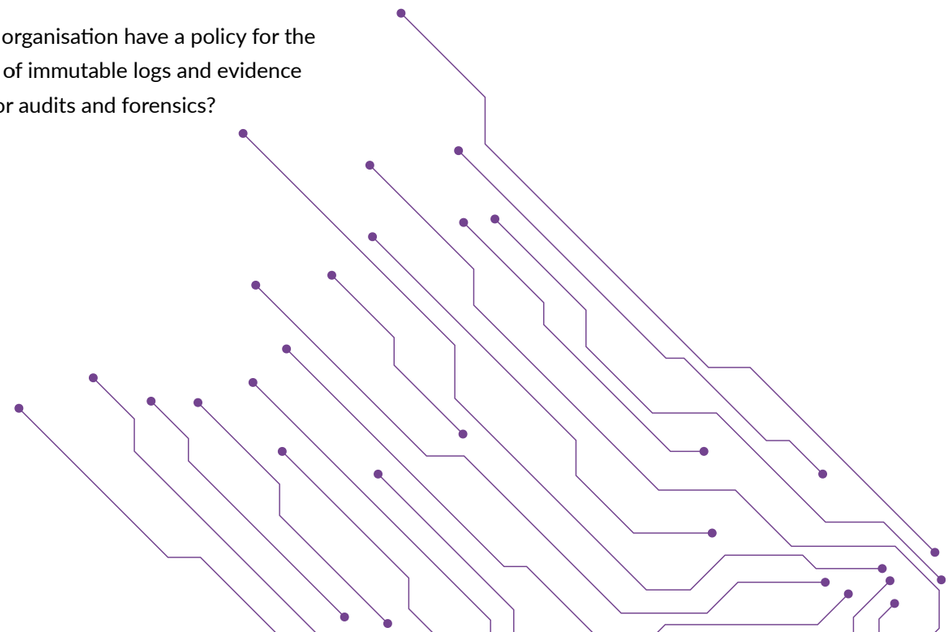
- Is there a system for ongoing monitoring of performance that detects drift, anomalous query patterns, and outputs that fall outside expected behaviour, with defined alert thresholds and alerting rules?
- Are regular validation tests scheduled against criteria established prior to deployment, so that degradation over time can be identified and remediated?
- Are mechanisms in place to collect, analyse, and act upon end-user feedback regarding inappropriate or unexpected model outputs?
- Is there a documented governance arrangement for retraining and redeployment outlining when retraining would occur, the methodology to use and the required approval process?
- Are monitoring and discovery tools active to detect the use of approved systems outside their agreed scope, prohibited use cases, unauthorised AI deployments and any Shadow AI?

Vulnerability and change management

- Do the developers provide security updates and patches and notify system operators and end users where appropriate?
- Are these security patches and updates tested and evaluated in a separate environment prior to deployment to ensure intended functionality?
- Where patching is not possible, is there a process in place to document this, communicate it and mitigate the risk through compensating controls?
- Is there a documented process to retest and assure AI behaviour whenever changes are made, such as major patches, model updates or modifications to system prompts?
- Are third party dependencies and external model components retested on a regular cadence or when the dependency itself is updated?

Incident response

- Are incident response playbooks, including rollback procedures and kill switches, available and checked regularly and have they been adapted to cover AI-specific risks?
- Does the organisation playbook align with the incident response runbook agreed with developers during the deployment phase?
- Has the organisation implemented logging of system and user actions sufficient to reconstruct incident timelines?
- Does the organisation have a policy for the retention of immutable logs and evidence needed for audits and forensics?
- Are regulatory notifications that may apply to an AI-incident including the EU AI Act, GDPR and NIS2 identified and built into the playbook?
- Is there a post incident review process to capture lessons learned, and is this feedback used to improve future design, testing and monitoring?



5. End-of-Life

Decommissioning AI solutions should be done proactively and responsibly to ensure that training data, model weights, configurations, and any derived artefacts are disposed of securely. This requires involvement and collaboration with key stakeholders, including data owners, who should be actively involved in the deletion or transfer of data.

Principle 7: Safely retire artefacts and document that compliance obligations have been met

The end of the AI lifecycle carries unique risks, particularly regarding the residual data which can be stored within model weights or training sets. Processes should include the systematic removal of access, the secure deletion or archiving of artefacts, and the preservation of an audit trail. The final stage of the AI lifecycle must ensure that the withdrawal of an AI system does not create new security vulnerabilities, and that the agreed exit strategy can be executed, reducing the risk of vendor lock-in. By conducting a final risk assessment and following a formal decommissioning plan, the organisation ensures that sensitive data is not orphaned and that all legal and/or contractual obligations are met.

Key Considerations

Decommissioning

- Is there a decommissioning plan in place with specified data deletion or archival procedures along with notification to affected users before proceeding with deletion of data or removal of access?
- Do relevant contracts and Service Level Agreements (SLAs) reflect termination obligations for suppliers?
- Has knowledge transfer been completed before decommissioning, ensuring that lessons learned are documented for incorporation into future projects?
- Does the decommissioning plan cover secure disposal of AI-specific artefacts (model weights, outputs, training and validation datasets, system prompts, embedding and vector stores) with the method for disposal documented?
- Have all credentials, API keys, service accounts, plugin registrations etc. been identified and revoked as part of decommissioning?

Documentation and Verification

- Has a final review of the decommissioned system been carried out documenting exactly what was deleted, when, and by whom to provide a clean audit trail?
- Does the organisation have a policy and process that documents what needs to be done when an AI system is retired, including the evidentiary records needed to demonstrate compliance?
- Have third parties documented confirmation of removal of artefacts from their systems?
- Has a residual risk assessment been completed before disposal with any lessons learned?

Summary

These guidelines translate international best practice, specifically ETSI EN 304 223, into a format that can be adopted by public sector bodies of varying sizes and integrated with the existing DPER Guidelines for the Responsible Use of AI in the Public Service. They are designed to be read alongside the NCSC AI Cyber Security Risk Assessment, which establishes the risk profile these guidelines are intended to address.

By adopting a secure by design approach from the outset, the appetite to deploy innovative solutions can be matched with an appropriate level of security and resilience. This approach ensures that sensitive public data is protected and systemic weaknesses are prevented before they emerge.

Although risks associated with AI continue to evolve, these principle-based guidelines which are aligned with the CyFun framework and ETSI standards provide the approach required to navigate these challenges safely.

Embedding these practices will ensure that AI in the public sector is deployed in a manner that is not only innovative but fundamentally resilient, secure and capable of meeting evolving challenges. To maintain this level of protection, these guidelines and the associated resources will be reviewed and updated on a regular basis to reflect the changing threat environment.

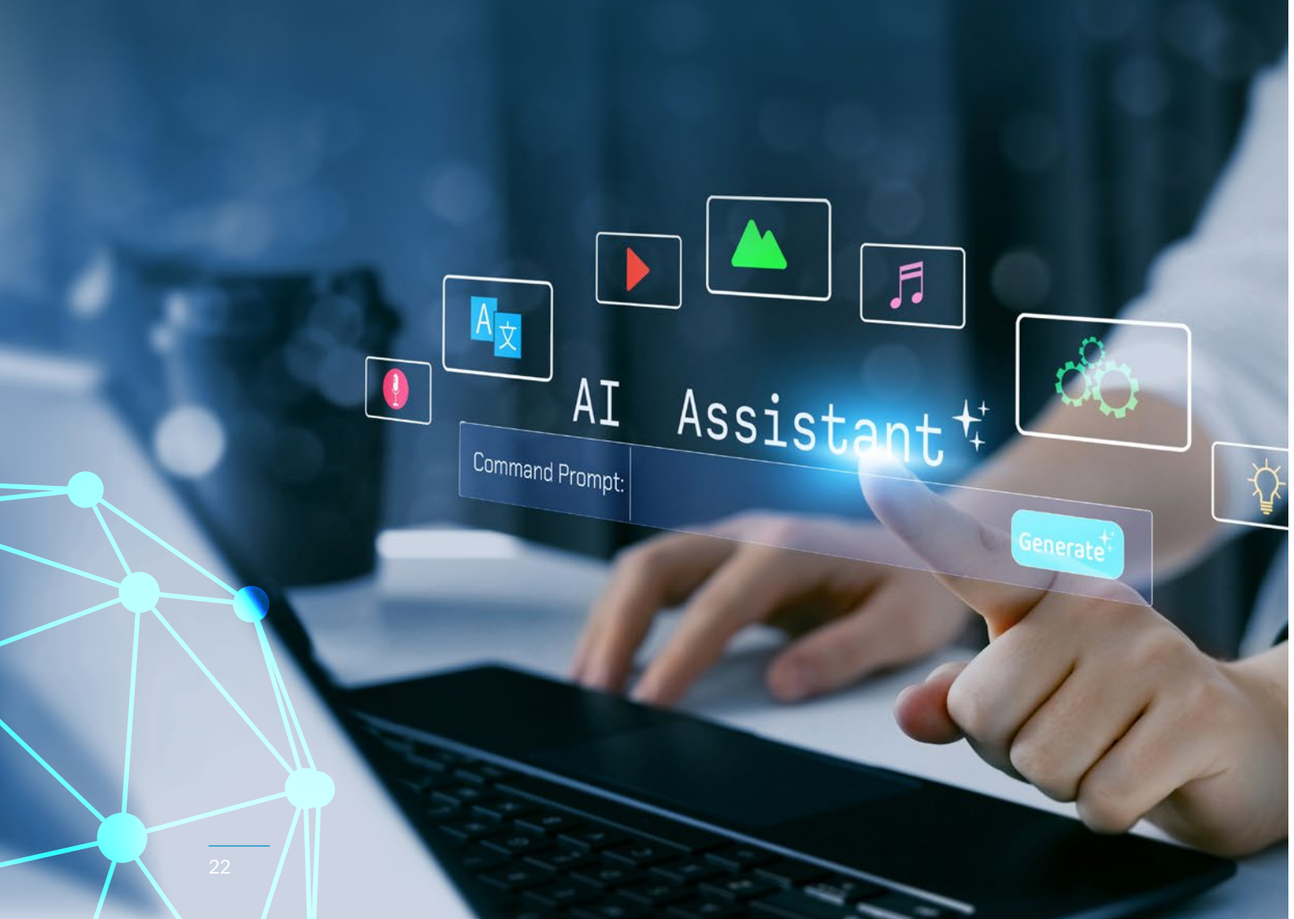
“Embedding these practices will ensure that AI in the public sector is deployed in a manner that is not only innovative but fundamentally resilient, secure and capable of meeting evolving challenges.”

Implementation Scenarios

The following scenarios illustrate how these guidelines would be applied in practice. They are worked examples, not precise templates. Each scenario shows how a public sector body could reason its way through a deployment and arrive at a proportionate set of controls, rather than a checklist to be reproduced step by step.

The two scenarios are drawn at deliberately different points on the spectrum of complexity and risk. The first is a common use case an organisation may encounter, a vendor-managed assistant used in a deliberate, user-initiated way. The second is more complex, a system the organisation builds and operates itself, exposed to the public as well as to internal staff. The level of control in each follows from the characteristics of the deployment, such as, how much of the system the organisation has built rather than bought, how broad its data access is, whether it is exposed to the public and therefore to adversarial use, and how consequential its outputs are.

The scenarios illustrate the point that these guidelines are flexible and can be applied in an appropriate and proportionate way, commensurate with the context and risk profile of each actual deployment.



Scenario 1:

A Department enabling a browser-based AI productivity assistant for user-initiated document work

A Department of approximately 500 staff decides to enable a browser-based AI productivity assistant. Staff access the assistant through a dedicated web interface in their browser and use it in a deliberate, user-initiated way: typing prompts for drafting help, ideas, or analysis, and uploading or pasting documents when they want help with a particular piece of content. The assistant does not have access to staff mailboxes, calendars, or files across the wider tenant. The user decides what to submit and when, and the assistant works only with what is submitted. This is the simplest deployment profile in the framework, and the cyber security work follows that profile.

Design

The team works through the DPER process first. The Decision Framework concludes the use is appropriate as a productivity assistant. The Responsible AI Canvas is completed setting out the high-level scope of the project and answering key considerations on important areas such as Human Agency and Oversight, Privacy and Data Governance and Legal Compliance and Oversight. A Data Protection Impact Assessment (DPIA) is completed along with assigning the appropriate EU AI Act risk classification.

The cyber security work focuses on an appropriately scoped set of decisions. An acceptable use policy is required and drafted, defining what staff may submit to the assistant: public information, working drafts, non-sensitive content, information that is not security classified. It prohibits submission of case files, personal data, security classified information, and any material covered by confidentiality undertakings. Access to the assistant is gated through the Department's Single Sign-On with mandatory multi-factor authentication, in line with the existing standard for any cloud-hosted service handling work content.

The vendor due diligence is appropriate and proportionate. The configuration is reviewed to confirm appropriate data residency, and that submitted content is not used for model training. The vendor's security assurance is checked to ensure that they have current security certificates such as CyFun, ISO 27001 or SOC 2 Type II and the reports are reviewed. Any AI-specific security documentation the vendor publishes is examined, and the vendor's sub-processor disclosure, data retention, and security incident notification commitments are confirmed against the Department's risk appetite. The ICT Steering Committee signs off on the project.

Development

There is limited development as the AI system is a managed feature accessed through the browser. Shared responsibility Threat Modelling is completed. Where the platform supports it, tenant-level controls are configured to restrict the file types and sensitivity levels that can be uploaded. The configuration applied is recorded as a single page in the Department's security baseline.

Deployment

A mandatory training module precedes access. It covers the acceptable use policy, the limitations of the assistant and the risk of over-reliance, the obligation to verify outputs against authoritative sources, and how the vendor handles content submitted to the service. It also covers the risks related to document upload: that the model processes everything in an uploaded file including metadata, tracked changes, hidden text, and embedded content, and that documents received from external parties may carry indirect prompt injection. Staff are instructed to upload only documents whose full contents they have inspected. Access is enabled only on completion of training.

A small pilot cohort uses the assistant for four weeks before broader rollout, which gives the Department a basis to confirm the assistant performs acceptably and the controls work as intended before scaling. The vendor's native audit logging is enabled at the highest available granularity, capturing user identity, timestamp, prompt content, output content, and references to any uploaded files, and retained in line with the Department's existing audit log policy. The retention period and log scope are checked against the vendor's documentation before rollout.

A defined reporting channel, through the existing IT service desk, is published to staff for raising concerns about outputs they believe to be incorrect, inappropriate, or indicative of a policy breach or security issue. The Department's coordinated vulnerability disclosure policy is updated to note whether it covers AI services, with reports triaged by the security team and escalated to the vendor where the issue lies in the service itself.

Maintenance

The acceptable use policy and the training content are reviewed annually and refreshed to reflect changes in the threat landscape and in the assistant's capabilities. The audit log is reviewed monthly by the security team for unusual use patterns or signs of policy breach.

Tenant-level discovery for unsanctioned AI tool use is enabled and reviewed monthly; the purpose of sanctioning this assistant is partly to reduce Shadow AI, and the monitoring verifies that staff are moving to the sanctioned tool rather than using it alongside unauthorised services.

The sanctioned assistant's own use is monitored for misuse patterns, signs that staff are submitting content the acceptable use policy prohibits, which are addressed through retraining, policy clarification, or tighter upload controls. A monthly verified check of the settings confirms that the "Opt-out of model training" configuration remains active and has not been reset by any updates.

The Department's existing incident response processes are confirmed to cover an incident involving the assistant, for example, a confirmed submission of sensitive content, or a vendor-notified security incident affecting the service. The regulatory notification paths under the EU AI Act, GDPR, and NIS2 are mapped against the plausible incident types, so that if an incident occurs the obligation to notify is already understood. When the vendor releases significant updates that change the assistant's behaviour, these are tested and once evaluated and approved, internal communication summarises the change for users.

End-of-Life

When the Department discontinues the assistant, licences are removed and account access is revoked, staff are notified, and the acceptable use policy is updated. The vendor is required, under the existing terms, to confirm deletion of stored prompts and outputs after the contracted retention period. A short residual risk assessment confirms that no business process has come to depend on the assistant in a way that would create operational risk on its removal, and a brief record documents what was removed and when.

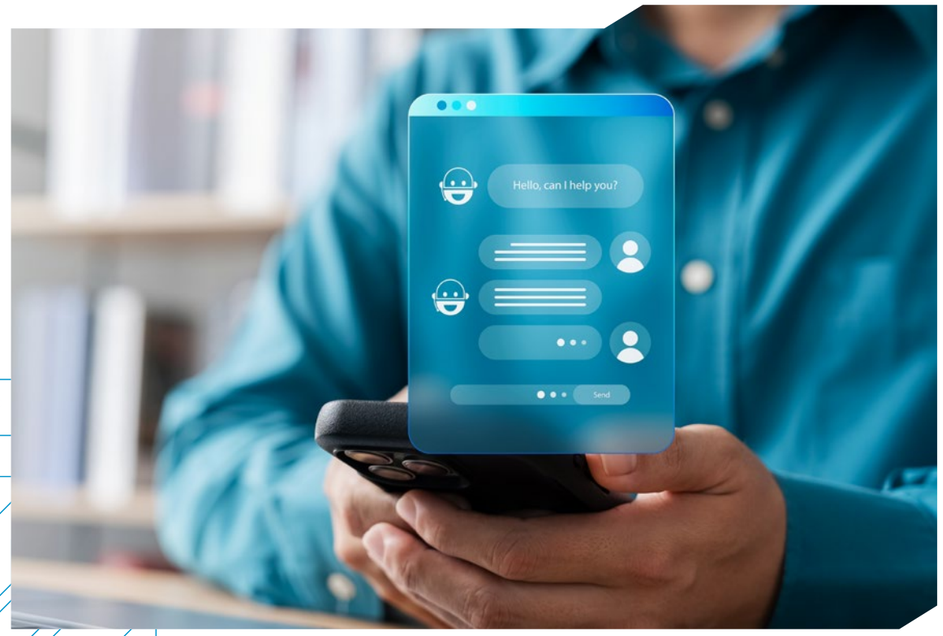
Scenario 2:

A Department building a public-facing AI chatbot for a citizen information service

A large Department running a high-volume citizen service, handling several hundred thousand public enquiries a year about entitlements, applications, and obligations, decides to build an AI assistant with two functions.

A public-facing version answers citizens' general questions and signposts them to the right process. An internal version, drawing on the same knowledge base, helps caseworkers find and apply policy. The assistant informs and signposts, however it does not determine eligibility or anything else affecting a person's rights, and that boundary is fixed in the design.

The Department decides to build the system itself and host it on a private cloud using an open-source foundation model, a deliberate choice to retain control of citizen query data and to build lasting internal capability.



Design

The team works through the DPER process first; the Decision Framework, the Responsible AI Canvas, the EU AI Act risk classification, and the DPIA. Combined these establish the foundation for responsible AI. The DPIA addresses citizen query content, interaction logs, and the handling of any personal data that a member of the public might enter.

The Project Steering Group, chaired by the Department CIO, with the Head of ICT Security, the Data Protection Officer, and the relevant Heads of Unit, owns approval. Its design decisions are recorded: to build the system, to host it on a sovereign private cloud, and to use an open-source foundation model selected against the Department's criteria for licence, provenance, maintenance status, accuracy on the relevant material, and the availability of security advisories. The selection decision also records what would trigger a change of model. The Department assigns a dedicated engineering team and engages a specialist delivery partner under a framework contract with a security schedule and knowledge-transfer obligations.

The architecture is designed around the public exposure from the outset. The public-facing tier is separated from the internal tier and from the core inference, orchestration, and data layers by distinct security zones, so that a compromise of the internet-facing component cannot reach the knowledge base or the internal system. Two data scopes are defined: the public assistant draws only on information cleared for public release, while the internal assistant can also surface internal guidance. Corpus scoping is carried out with the relevant data owners; live case files and personal data are excluded from both scopes.

Awareness is addressed from the start. The engineering team, the delivery partner, data owners, the caseworkers who will use the internal version, and the Steering Group are all briefed on the security implications appropriate to their role.

Development

The engineering team and the delivery partner's engineers complete AI-specific secure development training before development begins.

A threat model is completed which covers:

- the internet-facing application and its abuse and denial-of-service surface,
- adversarial users attempting jailbreaks,
- prompt injection, and extraction of internal-only content through the public interface,
- the inference infrastructure,
- the model weights as a sensitive asset; the document ingestion pipeline,
- the vector store,
- the identity layer, and
- the integrations with the source systems.

The infrastructure is built in line with the CyFun framework approach with security controls at every layer, including:

- encryption of model weights and embeddings at rest,
- network segmentation isolating the public tier,
- least-privilege service accounts,
- peer-reviewed infrastructure-as-code with change logging, and
- minimal hardened container images with vulnerability scanning.

- The public-facing tier sits behind rate limiting, abuse detection, and denial-of-service protection.
- External models and libraries are admitted only from an approved allow-list, with versions recorded.

The document ingestion and embedding pipeline is secured against poisoning, as the route by which the corpus reaches the model. Documents are validated and sanitised on ingestion to strip content that could act as indirect prompt injection, checked against the source system for provenance, and recorded in an immutable log. Only authorised owners can add to either corpus, any additions are reviewed.

Guardrails for the public are put in place including:

- every answer cites or links to its source,
- the model refuses out-of-scope requests,
- output filtering screens for personal data and unsafe content, and
- the public interface states at the point of use that the user is interacting with an AI assistant providing general information rather than a formal determination.

Internal users authenticate through Single Sign-On with multi-factor authentication and role-based access; the public interface is unauthenticated and constrained accordingly.

Functional accuracy is tested against a canonical question set; adversarial testing covers jailbreaks, direct and indirect prompt injection, and attempts to reach internal content through the public route; content-safety testing checks behaviour on hostile and sensitive inputs; and the infrastructure is penetration-tested across the public tier, the inference and orchestration layers, the ML Ops pipeline, and the vector store. Independent third-party red teaming covers the model and the infrastructure, with particular attention to the public interface.

The model and its environment are documented including :

- a model card with provenance and limitations,
- a dependency tree,
- an SBOM-equivalent for the open-source supply chain, and
- recovery procedures including model-weight restoration from secure backup.

Asset records track every component with provenance, version, and review cadence.

Deployment

Rollout is staged to reflect the public exposure. The internal version is piloted first with a cohort of caseworkers; the public version follows. Each gate is approved by the Steering Group.

Internal users complete training before access, covering the assistant's limitations, the duty to verify outputs against cited sources, and prohibited uses. The human oversight model is explicit: caseworkers remain the decision-makers, the public assistant signposts to cited public policies or human channels rather than resolving matters itself, and named operational personnel hold the access and authority to review the system's behaviour, suspend it, or roll it back. The public interface meets the Department's accessibility standard and carries the AI transparency notice.

A kill switch is configured and tested for each tier. The coordinated vulnerability disclosure policy is updated to cover the system, with a path for external researchers to report issues. Operational coordination with the delivery partner is documented; named contacts, escalation points, and a runbook for cooperating during a security or infrastructure incident.

Maintenance

The Department owns the whole system and therefore monitoring reflects that. Operational monitoring covers output quality and drift, user feedback is acted upon, and infrastructure health, latency, throughput, utilisation, error rates, and availability. The public tier is additionally monitored for abuse patterns, jailbreak attempts, and content-safety events, with thresholds that trigger review or intervention. Internal use is monitored for out-of-scope use, and tenant-level discovery for unsanctioned AI tools is enabled so the Department can see whether the sanctioned internal assistant is displacing Shadow AI.

Vulnerability management is also the Department's responsibility. The team tracks releases and advisories for the model, the inference and serving stack, and all dependencies; disclosed vulnerabilities trigger assessment, prioritisation, and remediation, with compensating controls documented and approved where prompt patching is not possible. Dependencies and external components are retested on a cadence and when updated. Model lifecycle management is structured: upstream releases that warrant an upgrade are evaluated and tested in staging before promotion, and a contingency plan covers model deprecation, or a critical vulnerability that is not patched, with a defined migration path.

Incident response playbooks cover AI-specific events (output anomalies, prompt injection, content-safety incidents on the public tier), infrastructure events (compromise, weight-exfiltration attempts, pipeline integrity failures), and the reputational dimension of a poor output reaching citizens. The playbooks align with the delivery partner's runbook. Logging is sufficient to reconstruct incident timelines and is retained per the Department's cyber security policies. Regulatory notification paths under the EU AI Act, GDPR, and NIS2 are mapped to event types. Every incident is followed by a post-incident review feeding back into design, testing, and monitoring. The infrastructure is penetration-tested annually.

End-of-Life

When the system is retired or replaced, the decommissioning plan covers each class of owned artefact, model weights, embeddings, vector store contents, prompts, and logs, with a specified deletion method and verification. The public are notified in advance and directed to alternative channels; internal users are supported through the transition; and knowledge transfer and lessons learned are documented for any successor. The infrastructure is decommissioned through the provider with confirmed data destruction, and all credentials, service accounts, and integration points are revoked. A final review records what was deleted, when, and by whom, and the Steering Group signs off the residual risk assessment.

Appendix I

Suggested Resources

Formal Standards

- ETSI EN 104 223 "Securing Artificial Intelligence (SAI); Guide to Cyber Security for AI Models and Systems" was published in May 2025. It breaks down requirements into five lifecycle phases to ensure models and systems are resilient against evolving threats The ETSI TR 104 128 "Securing Artificial Intelligence (SAI); Guide to Cyber Security for AI Models and Systems" provides further practical advice for implementing security across the AI lifecycle. This includes use cases and the application of the principles to them. See Appendix II for further details.
- The upcoming release of "of Artificial Intelligence - Cybersecurity Specifications for AI Systems" developed by CEN-CENELEC Joint Technical Committee 21 (JTC 21), is designed to provide a framework to comply with the robustness and cyber security requirements of the EU AI Act.
- ISO/IEC 42001:2023 standard is a globally recognised benchmark for AI management systems. It provides high level guidance on governance, outlining the controls and documentation needed to ensure that AI is designed and managed securely. The National Standards Authority of Ireland (NSAI) serves as the national member body for ISO and provides an outline of the standard.
- The ISO/IEC 27040 "Information technology — Security techniques — Storage security" provides further information related to the secure disposal of AI systems and the protection of underlying storage.

Irish Public Sector Guidance

- In 2025, DPER published the Responsible AI Canvas to facilitate teams in the initial design stage of an AI project. This provides a useful starting point to document the business case, risks and mitigations, human agency and oversight, accountability and compliance considerations. Further details on training and resources are available through DPER's AI Resources Artificial Intelligence Guidelines and Resources for the Irish Public Service. Their AI Lifecycle Guidance Tool includes considerations for the stages of an AI project within the Irish Public Service.
- The Office of Government Procurement released their guidance on Cloud Services Procurement Guidance Note in 2025, which contains a guidance note on procurement of cloud based AI services. Further guidance is expected under Ireland's forthcoming National Public Procurement Strategy which will consider the adoption of emerging technology such as AI.

- The Guidelines on Cyber Security Specifications published by the NCSC Ireland (NCSC-IE) provide a structured framework for public service bodies to embed cyber resilience into their ICT procurement. The guidelines are a valuable resource for AI security as they address supply chain security, data protection, and vendor attestation across the entire project lifecycle.

International Guidance

- The Department for Science, Innovation and Technology (UK) released their AI Cyber Security Code of Practice including the Implementation Guide for the AI Cyber Security Code of Practice in 2025. This provides principles, detailed case studies and practical scenarios to help organisations implement security provisions during model deployment. This guide applies the voluntary Code's requirements, outlining practical steps for real-world deployment environments.
- The NCSC UK published their guidelines on secure AI system development and provides a framework for building resilience into the AI lifecycle. They promote a "Secure by Design" mindset, outlining the importance of understanding the threat landscape and making system choices that are inherently resilient.
- The NCSC UK's Machine learning principles provides granular technical guidance on implementing secure Machine Learning models. It defines behaviour to address issues, such as data or model drift.
- The Cyber Security Agency of Singapore released their Guidelines and Companion Guide on Securing AI Systems. This adopts a secure by design approach. They include a number of detailed case studies including how to develop and implement a risk assessment.
- Two documents—ENISA's Framework for AI Cybersecurity Practices (FAICP) and ENISA's multilayer AI security framework (AI Good Cybersecurity Practices), provide comprehensive strategies for securing AI systems throughout their lifecycle.
- The EU Commission Public Buyers Community of practice published the Updated EU AI Model Contractual Clauses in 2024. These are included to provide a high-level description of requirements that must be met in commercial agreements, focusing on recommendations within the EU AI Act on Article 8 regarding accuracy, robustness, and cyber security measures.
- The NIST AI Risk Management Framework published by the National Institute of Standards and Technology (NIST) is designed to help organisations manage the risks associated with AI. Recently NIST released a Cybersecurity Framework Profile for Artificial Intelligence and further publications are expected.

Threat Landscape and Threat Modelling

- The Open Worldwide Application Security Project (OWASP) offers essential resources through the AI Exchange and the Gen AI Security Project. These provide a comprehensive and regularly updated view of the current AI security landscape, offering a strong starting point for identifying tools and recommendations based on emerging threats. Reviewing the latest AI Security Solution Landscape provides a good starting point of the current tools available and the recommendations.
- It is also important to remain informed by keeping up to date with any new and emerging risks and developments. The OWASP Top 10 Risks and Mitigations for LLM and Gen AI Apps and the Top 10 for Agentic Applications for 2026 provide information on the most common risks in production environments.
- Securing the deployment pipeline against the rise in supply chain attacks also requires the implementation of strict dependency controls and artefact management. OWASP published their OWASP GenAI Data Security Risks & Mitigations 2026 which provides a focus on data security risks and mitigations related to LLMs, GenAI, and Agentic AI. Updated Resources are regularly released on OWASP Resources Library
- The MITRE ATLAS (Adversarial Threat Landscape for AI Systems) matrix provides an overview of tactics and techniques targeting AI. This knowledge base can be used to inform threat modelling and red teaming exercises.
- The European Union Agency for Cybersecurity (ENISA) publishes an ENISA threat landscape report annually. The 2025 report provides a comprehensive analysis of the European cyber threat ecosystem, based on a study of nearly 5,000 incidents recorded to mid-2025. It examines what threats exist, but also how the underlying models of attack are shifting.
- Private sector organisations regularly publish threat landscape reports that provide vital insights into evolving risks. Notable examples include the Google Cloud AI Threat Intelligence Blog and the Mandiant special report on AI risk and resilience. Additionally, Anthropic provides targeted research including their August 2025 report on Detecting and countering misuse of AI: August 2025. Monitoring these industry updates ensures that the organisation stays informed about the latest adversarial tactics and defensive strategies.

Appendix II

ETSI guidelines and mapping to CyFun

ETSI EN 104 223 "Securing Artificial Intelligence (SAI); Baseline Cyber Security Requirements for AI Models and Systems" was published in 2025 followed by the ETSI TR 104 128 Securing Artificial Intelligence (SAI):Guide to Cyber Security for AI Models and Systems Technical Report which was published in January 2026; The Technical Report (TR) was produced by ETSI's Technical Committee Securing Artificial Intelligence (SAI). The ETSI Technical committee are preparing additional associated standards and further resources which will be reviewed by the NCSC-IE.

The 13 principles and 72 actions from the ETSI AI guidelines and the 7 principles from the NCSC-IE guidance have been mapped to the CyFun framework. This exercise was undertaken to show the linkage between the CyFun control categories and the AI Principles. The table below provides an example of this mapping for a single principle; the NCSC will be releasing supporting tools and materials including the full mapping to support implementation. The mapping exercise demonstrates that existing cyber security frameworks such as CyFun may need additional AI-specific controls at the implementation layer to achieve the framework category and subcategory outcomes.

NCSC-IE Principle - key consideration	CyFun	ETSI principle	AI-specific extensions
1 - Training and awareness	PR.AT-01, PR.AT-02, GV.RR-04	1 - Raise awareness of AI security threats and risks	• Is there a requirement for all staff using AI tools to complete AI literacy training before access is granted, covering acceptable use, data handling, and reporting of concerns along with the risks of Shadow AI? • Are awareness arrangements reviewed and refreshed on a defined cadence to reflect the evolving threat landscape?

Please note: *Any examples, opinions, or suggested approaches shared here are illustrative in nature and are not a substitute for your own governance, assurance processes, or independent professional advice. Organisations remain solely responsible for their own decisions (including security decisions), compliance obligations, and risk management outcomes.*

Appendix III

Glossary of key terms

Key Term	Description
AI Lifecycle	Refers to the structured, iterative framework used to manage the lifespan of an AI system, from initial conception to retirement or decommissioning.
Artefacts	Artefacts are defined as the core components of the AI system, including models, model weights, training and fine-tuning datasets, evaluation results, system prompts, audit logs, and any associated metadata or configuration files required to reconstruct or audit the system.
Data Owners	The individual or role accountable for the governance, quality, security, and appropriate use of specific datasets within an organisation, including decisions about access, retention, and lifecycle.
Dependency Tree	A map of all software libraries, frameworks, and external code that the AI system relies on to function. See also SBOM.
End Users	Individuals who interact with or rely on outputs from the AI system, whether within the organisation or external (for example, members of the public).
Knowledge Base Corpus	The curated collection of internal or public documents that the AI system retrieves to inform its responses.
Model Card	A standard document providing the provenance, intended use, limitations, and performance characteristics of an AI model.
Model Drift	A decline in an AI system's performance over time. This can happen either when the statistical properties of incoming data change relative to the training data (data drift), or when the underlying relationship between inputs and outputs changes even if input data appears similar (concept drift). Both forms result in the model relying on outdated patterns and producing less reliable outputs.
Model Weights	The numerical parameters of a trained machine learning model encoding its learned behaviour.
SBOM	Software Bill of Materials: A detailed inventory of the software components within a system, including open-source libraries, frameworks, versions, and licences, together with their dependency relationships. For AI systems, this may be extended (sometimes called an AI-BOM or ML-BOM) to also document model provenance, training data sources, and dataset versioning.
Shadow AI	The use of unauthorised AI tools by employees or contractors for work purposes, without organisational approval, governance oversight, or security review.
Supply chain (AI)	The full set of external components, services, datasets, pretrained models, and providers that an AI system depends on or incorporates, whether sourced commercially, from open-source projects, or from third-parties.

Appendix IV

Bibliography

CyFun. CyberFundamentals Framework. https://cyfun.eu/en

Cyber Security Agency of Singapore. (2024, October 15).
Guidelines and companion guide on securing AI systems.
https://www.csa.gov.sg/resources/publications/guidelines-and-companion-guide-on-securing-ai-systems

Department for Science, Innovation and Technology. (n.d.).
Code of practice for the cyber security of AI. UK Government.
https://www.gov.uk/government/publications/ai-cyber-security-code-of-practice/code-of-practice-for-the-cyber-security-of-ai

Department of Taoiseach (2026): Digital Ireland – Connecting our People, Securing our Future
https://www.gov.ie/en/department-of-the-taoiseach/campaigns/digital-ireland-connecting-our-people-securing-our-future/

Department of Public Expenditure, NDP Delivery and Reform. (2025, November 7).
Digital Public Services Plan 2030. Government of Ireland.
https://assets.gov.ie/static/documents/9cf1d031/Digital_Public_Services_Plan_web_fe.pdf

Department of Public Expenditure, NDP Delivery and Reform. (2025, May 7).
Guidelines for the responsible use of AI in the public service. Government of Ireland.
https://assets.gov.ie/static/documents/09fe3ad4/Guidelines_for_the_Responsible_Use_of_AI_in_the_Public_Service_20250918.pdf

Department for Science, Innovation and Technology. (2024, May 15).
Cyber security risks to artificial intelligence. UK Government.
https://www.gov.uk/government/publications/research-on-the-cyber-security-of-ai/cyber-security-risks-to-artificial-intelligence

Department of Enterprise, Trade and Environment:
AI - Here for Good A National Artificial Intelligence Strategy for Ireland
https://enterprise.gov.ie/en/publications/publication-files/national-ai-strategy.pdf

Digital Transformation Agency. Technical standard for government's use of artificial intelligence: Introduction, scope and target audience. Australian Government.
https://www.digital.gov.au/policy/ai/AI-technical-standard/technical-standard-governments-use-artificial-intelligence-introduction-scope-and-target-audience

ENISA – The Framework for AI Cybersecurity Practices (FAICP)
https://www.faicp-framework.com/

ENISA - Multilayer Framework for Good Cybersecurity Practices for AI
https://www.enisa.europa.eu/publications/multilayer-framework-for-good-cybersecurity-practices-for-ai

ENISA. (2025, October 1). ENISA threat landscape 2025.
European Union Agency for Cybersecurity.
https://www.enisa.europa.eu/publications/enisa-threat-landscape-2025

ETSI. (2025-12). Securing artificial intelligence (SAI); Baseline cyber security requirements for AI models and systems (TS 104 223 V1.1.1).
https://www.etsi.org/deliver/etsi_ts/104200_104299/104223/01.01.01_60/ts_104223v010101p.pdf

ETSI TR 104 128 Securing Artificial Intelligence (SAI); Guide to Cyber Security for AI Models and Systems
https://www.etsi.org/deliver/etsi_tr/104100_104199/104128/01.01.01_60/tr_104128v010101p.pdf

ISO. (2022). ISO/IEC 23894:2022 Information technology, Artificial intelligence, Guidance on risk management. International Organization for Standardization.

ISO. (2023). ISO/IEC 42001 Artificial intelligence, Management system. International Organization for Standardization.

MITRE ATT&CK. ATLAS: Security and assurance of AI. https://attack.mitre.org

MITRE Corporation – https://atlas.mitre.org/techniques/AML.T0010

National Cyber Security Centre. (2023, November 27).
Guidelines for secure AI system development. NCSC UK.
https://www.ncsc.gov.uk/collection/guidelines-secure-ai-system-development

National Institute of Standards and Technology. (2025, December 16). Cybersecurity Framework Profile for Artificial Intelligence (Cyber AI Profile): NIST Community Profile. (NIST IR 8596, initial public draft). NIST. https://csrc.nist.gov/pubs/ir/8596/iprd

NIST – Artificial Intelligence Risk Management Framework (AI RMF 1.0)
https://www.nist.gov/itl/ai-risk-management-framework

OECD – https://www.oecd.org/en/topics/ai-principles.html

Owasp GenAI Security Project https://genai.owasp.org/

OWASP. (2025, July 27). Securing agentic applications guide 1.0.
OWASP GenAI Security Project.
https://genai.owasp.org/resource/securing-agentic-applications-guide-1-0

OWASP. (2025, November 9). OWASP Top 10 for agentic applications for 2026.
OWASP GenAI Security Project.
https://genai.owasp.org/resource/owasp-top-10-for-agentic-applications-for-2026

OWASP. (2025, November 17). OWASP Top 10 for LLM applications 2025. OWASP GenAI Security Project. https://genai.owasp.org/resource/owasp-top-10-for-llm-applications-2025

UNESCO https://www.unesco.org/en/artificial-intelligence/recommendation-ethics

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