Security Standard – Artificial Intelligence & Machine Learning (SS-040)

Chief Security Office

Date: 29/10/2026

Department for Work & Pensions

This Artificial Intelligence and Machine Learning Security Standard is part of a suite of standards, designed to promote consistency across the Department for Work and Pensions (DWP), and supplier base with regards to the implementation and management of security controls. For the purposes of this standard, the term DWP and Department are used interchangeably.

Technical security standards form part of the DWP Digital Blueprint, which is a living body of security principles, architectural patterns, code of practice, practices and radars, that aim to support Product Delivery Units (PDUs) and suppliers in delivering the DWP and HMG Digital Strategy. Security standards and policies considered appropriate for public viewing are published here: DWP procurement: security policies and standards

Technical security standards cross-refer to each other where needed, so can be confidently used together. They contain both mandatory and advisory elements, described in consistent language (see table below).

Table 1 - Terms

Term	Intention
must	denotes a requirement: a mandatory element.
should	denotes a recommendation: an advisory element.
may	denotes approval.
might	denotes a possibility.
can	denotes both capability and possibility.
is/are	denotes a description.

1. Contents
2. Revision History 4
3. Approval History 4
4. Compliance 4
5. Exceptions Process 4
6. Audience 5
7. Accessibility Requirements 5
8. Introduction 6
9. Purpose 7
10. Scope 8
11. Minimum Technical Security Measures 9
11.1 Artificial Intelligence (AI) and Machine Learning (ML) Governance 9
11.2 Authentication Requirements 14
11.3 Deployments and Mandated Practices 17
11.5 Data Protection Controls 24
11.6 Monitoring and Logging 27
11.7 Model Security & Supply Chain 30
11.8 Evaluation, Red-Teaming, Assurance and Incident Response 32
11.9 ATRS Registration & Lifecycle Controls 35
11.10 Retrieval-Augmented Generation (RAG) Security 37
11.11 Secure End-Of-Life and Decommissioning 40
12 Appendices 41
Appendix A Security Outcomes 41
Appendix B Internal References 46
Appendix C External References 47
Appendix D Abbreviations 48
Appendix E Definition of Terms 48
Appendix F Accessibility artefacts 50
Appendix G AI-Specific Threat Examples (for Threat Modelling) 51

2. Revision History

Version	Author	Description	Date
1.0		First published version	29/10/2026

3. Approval History

Version	Name	Role	Date
1.0		Chief Security Officer	29/10/2026

This document is continually reviewed to ensure it is updated in line with risk, business requirements, and technology changes, and will be updated at least every 2 years - the current published version remains valid until superseded.

4. Compliance

Compliance with this standard will be verified through various methods, including but not limited to;

- controls tests performed by 1st line teams and by 2nd line activities (e.g. security testing teams)
- security assurance activities to ensure that Architectural Design and delivery are appropriate and aligned to applicable DWP Security Standards. [See Security Assurance Strategy – Ref. N].
- independent external audit
- Verification of Algorithmic Transparency Recording Standard (ATRS) registrations for all AI/ML components and services.

Results of these will be fed back to the appropriate DWP Risk and System Owners.

5. Exceptions Process

(Important) this section contains 'must' activities.

In this document the term **"must"** is used in bold letters to indicate a mandatory security measure. Any exceptions to the application of this standard, or where specific security measures cannot be adhered to, **must** be presented to an assigned DWP Security Architect, and considered for submission to the DWP Digital Design Authority (DDA) advisory or governance board where appropriate. This **must** be carried out prior to deployment and managed through the design caveats or exception process.

Such exception requests will invoke the Risk Management process to clarify the potential impact of any deviation to the configuration detailed in this standard.

Exceptions to the standard **must** be formally recorded by the Digital Design Authority (DDA). Additionally, System or Service Owners **must** maintain these exceptions on their local risk registers for accountability, traceability, and governance reporting.

6. Audience

This standard should be read by but not limited to technical architects, engineers, developers, security teams, Service and Risk Owners, and suppliers/third-party providers handling DWP data and services.

7. Accessibility Requirements

(Important) this section contains ‘must’ activities.
Users of this standard **must** consider accessibility design requirements as appropriate. Further information on accessibility standards can be found in Appendix F.

8. Introduction

(Important) this section contains ‘must’ activities.

This standard defines the minimum technical security measures required for AI/ML systems within DWP. It incorporates guidance from:

- NCSC Guidelines for Secure AI System Development (Secure design, development, deployment, and operation).
- DSIT AI Cybersecurity Code of Practice and Implementation Guide.
- NIST AI RMF v1.0 and Generative AI Profile
- NIST IR 8596: Cybersecurity Framework Profile for Artificial Intelligence
- NIST Special Publications:
    - SP800-53r5
    - SP800-161r1
    - SP800-218 (SSDF)
- DWP Internal Security Standards:
    - SS-001 - Parts 1 & 2 Access Control;
    - SS-002 - PKI and Key Management;
    - SS-003 - Software Development;
    - SS-007 - Use of Cryptography;
    - SS-011 - Containerisation;
    - SS-012 - Protective Monitoring;
    - SS-020 - Data Loss Prevention;
    - SS-027 - Security Testing;
    - SS-033 - Security Patching;
    - SS-036 - Secure Sanitisation and Destruction.
- ETSI TS 104 223 / TR 104 128.

The standard aims to:

- Ensure consistent implementation of AI/ML security controls.
- Mitigate risks from AI-specific threats (e.g., prompt injection, data poisoning, adversarial examples, model extraction, model inversion, excessive agency).
- Provide a baseline for compliance and assurance.
- Support achievement of NIST CSF-aligned security outcomes.

Governance
This security standard includes AI/ML-specific governance requirements such as including governance framework, risk classification and assessment, AI objectives, fairness and citizen impact assessment requirements, to support the secure implementation and assurance of AI/ML systems.

AI/ML systems **must not** execute autonomous, unverified actions where the output results in consequential citizen-affecting decisions or outcomes (specifically systems impacting benefit eligibility, entitlement calculations, fraud flagging, payment allocations, or legal rights) without human oversight. See section 11.4 for further information.

These governance requirements aim to address current governance arrangements which are emerging across the Department. As dedicated AI governance, frameworks and processes mature and are formally established, the relevant governance requirements within this standard will be repositioned to the appropriate governance artefacts. The governance requirements described in this document are mandatory – currently, any and all deployments of AI tools on DWP infrastructure **must** be approved by Digital Design Authority before deployment occurs

AI Risk Classification
The requirements within this standard **must** be applied using a risk-based approach. When utilising, developing, procuring, or deploying AI/ML systems, System Owners **must** understand and assess the risks associated with the system, taking account of factors such as data sensitivity, autonomy, citizen impact (including UK GDPR/DPA requirements), consequential decision-making, financial impact, and integration with Departmental systems.

The Department is establishing a formal AI/ML risk classification framework to support consistent assessment and proportional application of assurance, testing, monitoring, and governance requirements. Until this enterprise framework is formally ratified, technical risk thresholds **must** be assessed and controlled in accordance with the requirements of this standard.

9. Purpose

To ensure AI/ML systems are designed and operated safely and securely, aligned to DWP risk appetite, subject to consistent assurance, and resilient against AI-specific threats.

10. Scope

(Important) this section contains ‘must’ activities.
This standard applies to:

- All AI/ML models, agents, tools, datasets, pipelines, services, sandboxes, pilots, vendor trials and components processing DWP data;
- All suppliers using AI/ML within contractual boundaries;
- All third-party AI services integrated with DWP systems.

ATRS Compliance

All AI/ML systems **must** be assessed against the **Algorithmic Transparency Recording Standard (ATRS)** **before** production deployment (unless they meet exemptions criteria) and registration details maintained throughout the product lifecycle (see statement 11.9.1).

To support public-sector transparency the ATRS reporting template is split into two distinct tiers:
Tier 1 (Summary Information): Written in clear, simple, non-technical language aimed at the general public to explain what the AI tool does and why it is being used.

Tier 2 (Specialist Information): Split across multiple sheets containing granular technical specifications, data provenance, identified risks, and detailed mitigations aimed at specialist audiences (such as academics, journalists, and oversight bodies).

It **must** also be noted that where an AI/ML system processes clinical, medical, or health-related data, or is utilised to support health and disability assessments, the System Owner **must** consult the DWP Caldicott Guardian. The system **must** be formally evaluated to determine whether it qualifies as a medical device under Medicines and Healthcare products Regulatory Authority (MHRA) Software as a Medical Device (SaMD) regulations prior to deployment.

Any queries regarding the security measures laid out in this standard should be sent to an assigned DWP Security Architect or Project Team Lead.

11. Minimum Technical Security Measures

The measures below are the minimum required controls. Each item includes a mapping to NIST CSF sub-categories. (See Appendix A for a consolidated mapping.).

11.1 Artificial Intelligence (AI) and Machine Learning (ML) Governance

(Important) this table contains 'must' activities.

Reference	Minimum Technical Security Measures	NIST ID
11.1.1	Projects **must** document and adhere to the Departmental AI governance framework, including applicable governance forums such as Digital Transformation Group (DTG) and Responsible Artificial Intelligence Steering Group (RAISG), decision-making and approval authorities such as Digital Design Authority, assurance responsibilities, model ownership and risk ownership. See the Introduction section in this document for more information on current governance arrangements.	GV.RR-02 GV.PO-02
11.1.2	An AI Risk Assessment **must** be completed before deployment and maintained throughout the AI/ML lifecycle in line with NIST AI RMF (Appendix C).	ID.RA-04
11.1.3	AI/ML systems **must** undergo threat modelling proportionate to their risk, intended use, autonomy and deployment context. Threat modelling **must** include AI influenced offensive kill chain tactics such as automated reconnaissance, vulnerability research and exploit development (VRED) and privilege escalation, in addition to AI-specific threats, including prompt injection, data poisoning, adversarial examples, model extraction or inversion, excessive agency and misuse of connected tools.	ID.RA-04 ID.RA-05
11.1.4	Data lineage and model provenance **must** be documented (for example prompts, sources, licences, collection basis, quality and bias controls).	ID.AM-07

11.1.5	Auditability **must** be ensured for relevant lifecycle activities, including; • Logs for training, • Security and functional evaluation (in alignment with the AI Testing and Assurance Engineering Standard [Ref. S]) • model and prompt changes, • inference requests, and • agentic actions. This list is not exhaustive and may be expanded as necessary. If these requirements cannot be met, an exception **must** be requested from DWP Digital Design Authority.	PR.PS-04
11.1.6	Only essential AI functionality **must** be enabled based on the intended use and risk. Where advanced capabilities such as agent behaviour, tool execution, or automation are required, they **must** be explicitly justified and controlled.	PR.AA-05 PR.PS-06
11.1.7	Development, training, evaluation, and production environments **must** be separated logically or physically, dependent on risk.	PR.PS-01 PR.IR-01
11.1.8	Service capacity/resilience **must** be planned.	PR.IR-03
11.1.9	Model export or transfer outside approved environments, trust boundaries, or controlled lifecycle processes **must** be disabled unless explicitly approved.	PR.DS-10
11.1.10	Role-based access control **must** be applied to model management, tuning, and deployment workflows.	PR.AA-05

11.1.11	Documented AI/ML objectives **must** be established for each in-scope service or system. The assessment **must** confirm the use of AI/ML is appropriate, necessary, and proportionate for the intended purpose. The use of AI is applicable when the outcome cannot be achieved through an alternative without undermining business, security or user needs. The AI use **must** align with the Department's security expectations, data protection obligations, and risk appetite. Objectives and intended use **must** be measurable where practicable and reviewed when the model, data, supplier, operating environment, or use case materially changes.	ID.AM-08 ID.RA-07
11.1.12	For externally hosted, supplier-hosted, managed platform, or API-based AI/ML services, a documented AI Shared Responsibility Matrix **must** be established, approved, and maintained prior to live use. The matrix **must** define responsibility and accountability for: • Data controllership; • Model security; • Infrastructure security; • Identity and access management; • Data handling and retention; • Logging and monitoring; • Vulnerability management; • Patching; • Key and secret management; • Incident detection and response; • Backup and recovery; • Assurance evidence provision; • Secure change control; • Secure retirement and deletion of customer data; prompts; • Embeddings; • Indexes; • Model artifacts; • Skills libraries	GV.PO-01 GV.SC-02

11.1.13	Individuals involved in the design, development, deployment, operation, governance, assurance, or oversight of AI/ML systems, including but not limited to developers, technical architects, system operators, security teams, Service Owners, Risk Owners, and suppliers, **must** receive role-appropriate training regarding AI-specific security risks. Completion and review of this training **must** be evidenced as part of assurance activities.	PR.AT-02
11.1.14	AI/ML systems processing citizen data or supporting Departmental decision-making **must** utilise existing approved Departmental governance artefacts where appropriate, such as; • CDO Equality Assessment (which must evaluate algorithmic fairness, bias, and citizen impact) • DPIA (which **must** incorporate UK GDPR/DPA automated decision-making obligations as per the DWP Data Protection Policy [Ref. R]) Any identified gaps **must** be assessed and recorded in the AI assurance record. This assessment **must** be approved by the Digital Design Authority (DDA) and reviewed at least annually.	GV.OC-01 GV.OC-02 GV.OC-03
11.1.15	System owners **must** document and manage AI-specific 'technical debt'. This includes tracking outdated model dependencies, obsolete training libraries, and legacy prompt templates in the system documentation to prevent long-term security and operational vulnerabilities.	ID.AM-02 ID.AM-08
11.1.16	System owners **must** document, manage and mitigate AI-specific technical risk. To maintain appropriate independence from delivery teams and manage conflicts of interest, security assurance of AI/ML systems **must** be performed via the established Digital Security Risk Management (DSRM) and Enterprise Security Risk Management (ESRM) controls assessment and assurance processes as appropriate, supplemented by AI-specific threat modelling.	GV.RR-02

11.1.17	AI/ML systems, assistants and agentic capabilities **must** be approved, inventoried and managed through an accountable Departmental service or platform owner. Approved strategic Departmental platforms **must** be used where they meet the documented use case. Supplier-managed or SaaS-integrated AI capabilities **must** be subject to supplier assurance, documented shared-responsibility arrangements, approved configuration, logging and access controls. Deployment teams **must** consult the latest NCSC guidance on agentic deployments, such as 'Managing the cyber risk of agentic AI' [see External References] prior to deployment. The deployment or use of unapproved, unmanaged, or 'shadow' AI applications on local endpoints is strictly prohibited. Where endpoint-local AI execution is explicitly approved by exception, it **must** enforce security, logging, and access controls equivalent to centrally managed services.	PR.PS-04 PR.IR-01
11.1.18	Where multiple delivery projects or programmes utilise an identical supplier-hosted or third-party AI code analysis platform to inspect non-public or legacy source code, a single, comprehensive Enterprise Platform Environment Risk Assessment **must** be conducted, subject to formal SRO risk acceptance and verified technical guarantees that the supplier platform enforces memory-only execution and zero data retention for model training. Individual delivery projects **may** inherit this environment-level assurance via fast-track governance by correspondence, avoiding repetitive project-level platform re-assessments.	ID.RA-01

11.1.19	Digital Design Authority approved Departmental strategic AI platform services **must** be used where they provide the capability required by the approved use case. A separate or locally developed platform **must not** be adopted solely to duplicate an available strategic capability. Where the strategic platform cannot meet documented business, technical or security requirements, the alternative design **must** be assessed and approved through the applicable architectural and security governance routes (and approved by Digital Design Authority) before deployment.	PR.PS-01
11.1.20	Each AI service **must** maintain an AI assurance record containing governance, security, privacy, supplier, evaluation and testing evidence required to support risk decisions throughout the service delivery lifecycle. This record may include multiple artefacts.	ID.AM-02

11.2 Authentication Requirements

Please refer to SS-003 Software Development Security Standard [Ref. F] for more information on API Authentication requirements.
(Important) this table contains ‘must’ activities.

Reference	Minimum Technical Security Measures	NIST ID
11.2.1	AI/ML systems **must** define and document their authentication architecture before deployment. Federated identity mechanisms **must** be used where available; • Users, agents, services and APIs **must** authenticate using DDA approved identity services, • only authorised AI/ML services, models, tools and endpoints may be deployed in the execution environment. • microservices/gateways **must** authenticate to DDA approved services; • service registry **must** validate only legitimate AI/ML services.	PR.AA-03

11.2.2	Public-facing AI/ML services **must** use approved federation and token-based authentication mechanisms in accordance with SS-001 Part 1 Access & Authentication Security Standard [Ref. A] and SS-002 PKI & Key Management Standard [Ref. B]. Protocol and token profiles **must** follow current DWP-approved implementation patterns.	PR.AA-03
11.2.3	AI agents **must** operate using uniquely identifiable service or workload identities. Agentic systems **must not** share user credentials or operate using generic accounts in line with SS-001 Part 1 Access & Authentication Security Standard [Ref. A].	PR.AA-02
11.2.4	Token generation **must** use approved algorithms as per SS-007 Use of Cryptography Security Standard [Ref. C].	PR.DS-01
11.2.5	Where an AI agent/tool acts on behalf of a user, the delegated authority **must** be managed using approved delegated identity mechanisms and least privilege. Access control, authentication, delegated permissions, and session management **must** be implemented in accordance with SS-001 (Part 1): Access and Authentication Security Standard [Ref. A] and SS-001 (part 2): Privileged User Access Security Standard [Ref. P]. Agent actions **must not** rely on pooled, generic, or shared privileged identities in place of the invoking user's approved security context. Every downstream API execution **must** require OAuth Token Exchange to down-scope the agent's token to match the exact delegated permissions and session context.	PR.AA-02 PR.AA-05
11.2.6	AI systems **must** implement authentication, identity lifecycle management, access control and credential protection in accordance with SS-001: (Part 1): Access and Authentication Security Standard [Ref. A] and SS-002: Public Key Infrastructure & Key Management [Ref. B].	PR.AA-02 PR.AA-03

11.2.7	Access, identity and refresh tokens used by AI/ML services and agents **must** be validated and managed in accordance with SS-001 Part 1 Access and Authentication standard [Ref. A] and applicable DWP identity patterns. Expired, invalid, revoked, incorrectly scoped or incorrectly audience-bound tokens **must not** be accepted. Requests using such tokens **must** be rejected with a protocol-appropriate authentication failure response.	PR.AA-01 PR.AA-03
11.2.8	Expired tokens **must** be purged in line with SS-001 Part 1 Access and Authentication Security Standard [Ref. A].	PR.AA-01
11.2.9	Delegated privileged or consequential actions **must** be protected against unauthorised replay. Replay protection **must** be proportionate to the transaction risk and use approved controls such as audience and resource binding, short-lived credentials, transaction identifiers, and nonces where applicable.	PR.AA-02
11.2.10	Credentials used by AI/ML systems, agents, gateways and tool integrations **must** be generated, protected, stored, rotated, revoked and validated in accordance with SS-001 Part 1 Access & Authentication Security Standard [Ref. A], SS-002 PKI & Key Management Security Standard [Ref. B] and SS-007 Use of Cryptography Security Standard [Ref. C]. AI prompts, skills, model artefacts and orchestration configuration **must not** contain embedded credentials or secrets.	PR.AA-01 PR.AA-04
11.2.11	Developers **must** implement fine-grained authorisation at the control plane using the Model Context Protocol (MCP) with OAuth token exchange in line with requirements in SS-003 Software Development Security Standard [Ref. F]. MCP access control lists (ACLs) **must** restrict agents to executing only authorised methods and tools, preventing unauthorised write actions, file deletions, or lateral system access.	PR.AA-05

11.2.12	Agent authority **must** be documented and approved for systems capable of creating, modifying, approving or executing actions affecting production systems, claimant data, payments or Departmental records that require enhanced assurance and monitoring.	DE.CM-09

11.3 Deployments and Mandated Practices

(Important) this table contains ‘must’ activities.

Reference	Minimum Technical Security Measures	NIST ID
11.3.1	AI/ML services **must** operate in alignment with network segmentation and egress controls appropriate to their trust boundary and deployment model. Where supported by the platform, these controls **must** be implemented and maintained through approved automated policy mechanisms.	PR.IR-01
11.3.2	Underlying software, certificates, services, and infrastructure **must** be maintained in line with SS-033 Patching & Exposure Management Standard [Ref. D]. Internet facing AI-Gateways (or equivalent technical controls), model runtimes, orchestration components and Model Context Protocol (MCP) servers **must** be subject to continuous exposure management and prioritised vulnerability remediation. Where an exposed component cannot be remediated within the timescales required by SS-033 Patching & Exposure Management Standard [Ref. D], compensating controls **must** be implemented and evidenced. These may include isolation, removal of internet exposure, disabling affected functionality, access restriction, enhanced monitoring or temporary service suspension.	ID.AM-08 PR.PS-02

11.3.3	All traffic between AI/ML endpoints (including internal) **must** be encrypted in transit in line with SS-007 Use of Cryptography Security Standard [Ref. C].	PR.DS-02
11.3.4	AI/ML components **must** enforce execution within either: • The delegated user identity and permission context; or, • an explicitly approved and governed non-human identity context. All executions **must** remain attributable, auditable, and aligned to approved access control mechanisms.	PR.AA-01 PR.AA-02
11.3.5	CI/CD pipelines **must** be secure-by-default and align with the NIST Secure Software Development Framework (SSDF / SP 800-218) [Appendix. C].	PR.PS-06
11.3.6	Model artefacts and updates **must** be version-controlled and cryptographically signed, and hashes recorded in line with SS-002 PKI & Key Management Standard [Ref. B].	PR.DS-01 PR.PS-06
11.3.7	AI/ML systems **must** maintain a clear trust boundary between system or orchestration instructions and untrusted content, including user input, retrieved documents and external tool output. Untrusted content **must** be distinctly represented and handled so that it cannot modify trusted instructions, authorise actions or bypass independent access and policy controls.	PR.PS-05 PR.PS-06
11.3.8	Third-party AI/ML services **must** undergo a security review by ESRM Supply Chain Assurance prior to procurement and **must** be verified prior to Digital Design Authority (DDA) submission (see NIST SP 800-161) [Appendix C]. Supplier assurance **must** include provision of the documented AI Shared Responsibility Matrix and evidence that responsibilities have been assigned, accepted, and reviewed.	GV.SC-07

11.3.9	AI/ML systems **must** only be deployed and used in accordance with their documented intended use, approved operating constraints, and supporting documentation. Any proposed use outside documented conditions **must** be subject to reassessment and approval before implementation.	ID.AM-08 PR.PS-01
11.3.10	System prompts, tool instructions, agent policies, steering files, guardrails, retrieval instructions, and curated AI Skills Libraries (including third-party prompt packages, agent skill definitions, and instruction templates) **must** be treated as controlled configuration items and software supply-chain assets. They **must** be version-controlled, code-reviewed, hardened, verified, approved, access-controlled, vulnerability-scanned for embedded prompt injection vectors, and protected from unauthorised disclosure or modification. Secrets, credentials, tokens, connection strings, privileged configuration data, and other sensitive security information **must not** be embedded in these artefacts. System prompts and related orchestration artefacts **must not** be relied upon as the sole enforcement mechanism; independent technical controls **must** enforce access, validation, and policy compliance. Importing unvetted, third-party AI skills or prompt repositories directly into production workflows is strictly prohibited, and all AI skills **must** be validated against approved DWP instruction baselines before deployment.	PR.DS-01 PR.PS-01 PR.PS-06

11.3.11	To prevent unbounded resource consumption, AI deployment environments **must** implement a dedicated AI Gateway (or equivalent technical controls). The following gateway requirements **must** be considered, proportionate to risk; • Identity-Bound Token Rate Limiting per user, per agent session, and per model family; • Semantic Caching to deflect repetitive queries and reduce latency; • Automated Provider Failover to maintain system availability; • A deterministic, operator-accessible 'kill switch' to suspend execution immediately when defined thresholds or runaway loop behaviours are detected.	PR.IR-03 PR.IR-04
11.3.12	Secrets, credentials, API keys, signing material, and connection strings used by AI/ML services **must** be stored, issued, rotated, and revoked using approved DWP secrets-management and key-management mechanisms in line with SS-002 PKI & Key Management Standard [Ref. B]. Plaintext production secrets, passwords, or API keys **must not** be stored on local developer disks or within directories indexed by AI assistants.	PR.DS-01
11.3.13	AI/ML systems that execute code, terminal commands or package installation **must** run within an approved isolated execution environment (i.e. a sandbox) proportionate to risk. The environment **must** prevent privileged actions being performed outside of the sandbox, restrict filesystem and network access, prohibit unapproved host mounts and credentials, enforce resource limits and be disposable or recoverable. Any elevated capability within the isolated environment **must** be explicitly required, bounded to the approved task and **must not** confer equivalent privilege outside that environment. Developer containers (e.g. local Dev Containers) may be used as a transitionary recommended best practice but are not accepted as primary security boundaries in line with SS-011 Containerisation Security Standard [Ref. E].	PR.IR-01

11.3.14	All AI-generated outputs, completions, tool-call parameters, and intermediate reasoning traces **must** be treated as untrusted data until validated by independent technical controls or verified human review. Model outputs **must not** be directly executed as shell commands, interpreted as executable system code, or consumed by downstream APIs without passing automated structural schema validation, parameter sanitisation, and authenticated session down-scoping.	PR.PS-05 PR.PS-06
11.3.15	Agentic AI systems **must** operate within controlled execution environments that restrict access to networks, systems, tools, services and data to only those resources necessary to perform their approved functions.	PR.IR.01 PR.PS-05
11.3.16	Agentic AI systems **must** provide authorised personnel with the ability to rapidly suspend, isolate, disable or revoke agent execution where unexpected, unsafe or unauthorised behaviour is identified.	PR.PS-05 PR.IR-01

11.4 Non-Repudiation Safeguards
(Important) this table contains 'must' activities.

Reference	Minimum Technical Security Measures	NIST ID
11.4.1	For actions such as, but not limited to: • Payments; • Configuration changes; • Citizen-facing systems; • Consequential citizen-affecting outcomes; • Decisions which impact the Department's security or resilience capability. AI/ML systems **must not** execute autonomous, unverified actions where the output results in consequential citizen-affecting decisions or outcomes (specifically systems impacting benefit eligibility, entitlement calculations, fraud flagging, payment allocations, or legal rights). Human oversight requirements **must** be applied commensurate with risk.	PR.AA-05
11.4.2	Log attribution metadata **must** be generated for user requests and agent actions, with timestamps and correlation IDs.	ID.AM-07 PR.PS-04

11.4.3	Every agent-initiated privileged action **must** be cryptographically verifiable and enforce cryptographic non-repudiation, in line with SS-007 Use of Cryptography Security Standard [Ref. C] and SS-002 PKI & Key Management Security Standard [Ref. B], recording the originating user or workflow, relevant approval state, and traceability metadata.	PR.AA-05
11.4.4	All AI-driven activities - specifically decisions, automated completions, downstream system actions, code execution, shell command usage, and code commits - **must** be fully traceable. To achieve this, the system **must** record and bind the following metadata to the action: • **Model Context:** The specific AI model family, the version used, and the underlying prompt or instruction set. • **Identity Context:** The originating user, an approved non-human identity, the triggering workflow, or the cryptographically verified workload identity. • **Code Provenance:** For co-authored code, developers **must** self-declare this provenance within the repository using structured Git commit trailers (e.g., AI-Assisted, AI-Tool, AI-Model) in accordance with SS-003: Software Development Security Standard [Ref. F].	PR.PS-04
11.4.5	Actions such as but not limited to payments, configuration changes, code deployments, citizen-facing systems and consequential citizen-affecting outcomes **must** enforce cryptographic signing in line with SS-007 Use of Cryptography Security Standard [Ref. C] commensurate with risk.	PR.DS-02
11.4.6	AI/ML outputs **must** be subject to validation controls proportionate to risk, ensuring outputs are accurate, appropriate, and safe before driving decisions or actions. Where validation cannot be fully automated, appropriate human review or secondary verification mechanisms **must** be implemented.	PR.DS-10

11.4.7	AI-generated outputs, completions, recommendations, and intermediate reasoning traces **must** be treated as untrusted data until validated through independent technical controls or human review. Outputs **must not** be directly executed as code or consumed by downstream APIs without schema validation, parameter sanitisation, policy enforcement, and delegated session authorisation.	PR.PS-06
11.4.8	AI-generated, suggested or co-authored code **must not** be merged into a protected branch or released solely on the basis that it was generated by an approved tool. It **must** undergo the peer review, automated security scanning, dependency checking and testing required by SS-003: Software Development Security Standard [Ref. F]. An authorised person **must** review the purpose, security implications and relevant test evidence and accept the change through the approved repository workflow.	PR.PS-06
11.4.9	High-level AI co-authorship traceability **must** be self-declared at the git-tree level using structured commit trailers. When AI tooling materially contributes to a commit, developers **must** append the following metadata separators to the commit body: AI-Assisted: true, AI-Tool, and AI-Model.	PR.PS-06
11.4.10	Human oversight mechanisms **must** enable meaningful review and intervention. The oversight workflow **must** be specifically designed to mitigate 'automation bias' (the tendency to over-rely on automated outputs). Reviewers **must** possess appropriate authority, training, and sufficient contextual information to robustly challenge, verify, reject, or override AI-generated recommendations independently of the AI's own confidence scoring.	GV.RR-02 PR.AT-02

11.5 Data Protection Controls

Where an AI/ML system processes personal data, introduces AI into an existing process involving personal data, or uses AI-generated outputs in a way that affects individuals, the DWP Data Protection Impact Assessment process must be followed before the processing begins and the assessment must be reviewed where the processing, data, model, purpose or associated risks materially change.

(Important) this table contains ‘must’ activities.

Reference	Minimum Technical Security Measures	NIST ID
11.5.1	Data minimisation controls **must** be applied. For training, testing, and evaluation, synthetic, obfuscated, pseudonymised, or otherwise privacy-reduced data (which may still constitute personal data, and thus subject to applicable data protection requirements) **must** be used when the business purpose can be achieved, in line with the DWP Data Obfuscation Security Policy [Ref. I]. Where production or personal data is necessary, the legal basis, necessity, proportionality, and security controls **must** be documented and **must** have an approved exception in place.	PR.DS-01
11.5.2	Use of special category data for any AI or ML model training is prohibited unless explicitly authorised by the Data Owner, following completion of applicable data protection governance processes and formal risk acceptance by the Risk Owner.	PR.DS-01
11.5.3	AI systems and supporting gateways **must** implement preventive (proportionate to risk), detective controls to identify unauthorised submission, retention, memorisation or disclosure of personal and sensitive data. Controls may include input classification, data-loss-prevention rules, output monitoring, canary records, redaction and alerting.	DE.CM-09
11.5.4	Access to training datasets and feature stores **must** be restricted to authorised roles. Approvals **must** be enforced for all data exports.	PR.AA-05

11.5.5	Risk assessment **must** determine whether privacy-enhancing technologies are required. Where the assessment determines they are required, appropriate privacy-enhancing technologies **must** be implemented and the trade-offs documented.	ID.RA-01
11.5.6	Ingestion pipelines **must** authenticate source repositories and sign ingested documents; maintain tamper-evident indexes in line with SS-003: Software Development Security Standard [Ref. F].	PR.AA-03
11.5.7	Content sanitisation (HTML/Markdown/links/scripts) **must** be applied, and remote execution/rendering in downstream components **must** be disabled.	ID.AM-08
11.5.8	DWP data, prompts, outputs, feedback, and telemetry **must not** be used for provider-side training, tuning, or model improvement unless that use has been explicitly approved by the Data Owner, and **must** be; • lawfully authorised, • contractually bound, • reflected in the ATRS, and • risk assessed.	GV.SC-05

11.5.9	Processing data classified as OFFICIAL-SENSITIVE, personal claimant data, or non-public Departmental source code using unassured, consumer-grade public SaaS AI services is not permitted unless the conditions below are satisfied. Ingestion of sensitive DWP datasets into commercial foundation models or third-party AI services is permitted **only** where the service is deployed within an assured, tenant-isolated enterprise cloud boundary (governed under SS-023: Cloud Security Standard [Ref. Q] and SS-006: Security Boundaries Standard [Ref. M]), enforces zero data retention for provider-side model training, and is supported by an approved DPIA where applicable and risk assessment. Information subject to an applicable originator handling instruction, dissemination restriction which prohibits or restricts its disclosure, upload or processing by a non-governmental or third-party service **must not** be entered into AI/ML without documented authorisation from the originator or an authority formally empowered to vary the restriction. Tenant isolation, contractual controls, zero provider-side training and enterprise licensing **must not**, by themselves, be treated as overriding an applicable originator restriction.	PR.DS-01
11.5.10	When training, testing, or fine-tuning models in non-production environments, data teams **must** use synthetic, masked, or obfuscated data in compliance with the DWP Data Obfuscation Security Policy [Ref. I]. Where use of production data is required e.g. to conduct bias or fairness testing, an exception **must** be sought as per section 11.5.1. Non-production environments **must** maintain strict, logical separation from production networks.	PR.DS-01 PR.PS-01

11.6 Monitoring and Logging

(Important) this table contains ‘must’ activities.

Reference	Minimum Technical Security Measures	NIST ID
11.6.1	AI/ML services, gateways and agentic systems **must** produce logs in accordance with SS-012 Protective Monitoring Security Standard [Ref. J]. Logs **must** cover relevant AI/ML components and provide an attributable record of agent actions, including the initiating identity or workflow, instructions or skills used, tools invoked, actions taken and their outcomes. Logging **must** support protective monitoring, incident investigation and forensic investigation.	PR.PS-04
11.6.2	API and inference logs, including validation errors and crashes, **must** be logged in accordance with SS-012: Protective Monitoring Standard [Ref. J].	PR.PS-04
11.6.3	AI/ML monitoring capabilities and central AI Gateways (or equivalent technical controls) **must** actively detect, log, and alert on model extraction, systematic token harvesting, and side-channel probing attempts. Protective monitoring **must** inspect for abnormal query volumes, sequential prompt permutations, repetitive semantic boundary probing, abnormal vector database query churn, and observable inference side-channels (such as anomalous token latency or cache hit distributions). Upon detection of extraction activity, central AI Gateways **must** automatically enforce rate-limiting, challenge the active session, or suspend the invoking API credential.	PR.PS-04 DE.CM-09
11.6.4	Model-aware detection **must** be established for anomalous prompt patterns, abuse of tools, extraction attempts, and drift signals.	PR.PS-04 DE.CM-09

11.6.5	AI/ML services **must** have documented AI-specific detection use cases and threat-hunting procedures proportionate to risk. These **must** include, where relevant: • prompt injection and guardrail-evasion campaigns; • model extraction and systematic token-harvesting patterns; • repeated failed tool or MCP authorisation; • anomalous or repeated agent execution loops; • retrieval poisoning and unauthorised corpus modification; • abnormal embedding, vector-store or index activity; • material degradation in grounding or citation behaviour; • automated network, cloud-resource, identity or repository discovery; • abnormal credential access or credential-harvesting activity; • unauthorised privilege escalation; • lateral movement between connected tools, MCP servers, repositories, services or trust boundaries; • anomalous tool selection, tool-call sequences or execution outside the approved use case; and • indicators that AI/ML services, agents or models are being used to support command-and-control, exploitation, persistence, exfiltration or other cyberattack activity. Detection logic **must** correlate relevant identity, endpoint, network, cloud, AI gateway, model, tool-call and MCP telemetry where available. Alert thresholds, response procedures and ownership **must** be tested and reviewed following incidents, material architectural changes or significant changes to the threat landscape.	DE.CM-09

11.6.6	Protective monitoring and logging for agentic and RAG workflows **must** capture additional domain-specific metadata. Logging pipelines **must** record vector store query latency, embedding similarity scores, retrieved document identifiers, and Model Context Protocol (MCP) tool-call parameters to support post-incident forensics.	PR.PS-04

11.7 Model Security & Supply Chain

(Important) this table contains ‘must’ activities.

Reference	Minimum Technical Security Measures	NIST ID
11.7.1	Model weights and datasets **must** be verified via cryptographic hashes/signatures.	PR.DS-01
11.7.2	System owners **must** maintain an up-to-date AI Bill of Materials (AIBOM), Machine Learning Bill of Materials (ML-BOM), or equivalent formally governed assurance asset record. As a minimum, this record **must** identify and explicitly document: • model family and version, • provider, • weights or checkpoints where applicable, • fine-tuning sources and datasets, • embedding models, • vector store configurations, • orchestration frameworks, • prompt templates, • tool connectors, • guardrail components, • inference endpoints, • hosting region (where available), • foundation model provider, • model release date, • external tool dependencies, • MCP servers, • Retrieval sources, • External AI Skills libraries used and their sources, • and all external prompt templates, and agent tool-definition packages utilised by the system, including their relevant licenses, source provenance evidence, and cryptographic hashes / integrity verification data.	ID.AM-02 ID.AM-07

11.7.3	Third-party and open-weight models **must** be evaluated before adoption for; • licence terms and intended-use restrictions, • country of origin and vendor affiliation, • provenance and integrity, • known vulnerabilities, • security advisories, • malicious artefacts, • embedded secrets, • unsafe dependencies, • operational constraints. Evidence of evaluation **must** be retained in the system assurance record and referenced in the ATRS entry where applicable.	GV.SC-07
11.7.4	Third-party model files, open-weight models, checkpoints, serialised weights, and associated artefacts **must** be treated as untrusted third-party code until verified. AI/ML systems **must not** load insecure, arbitrary-code-execution-prone serialisation formats (such as Python's legacy Pickle utility). Model execution runtimes **must** enforce safe serialisation formats (such as Safetensors). Imported model artefacts **must** be scanned, integrity-checked, sandboxed where appropriate, and executed only within approved isolated environments.	ID.RA-01 ID.RA-09
11.7.5	Model runtime and tools **must** be isolated.	ID.RA-01
11.7.6	File system and network egress for AI/ML models **must** be restricted.	PR.IR-01
11.7.7	AI/ML tool allowlisting **must** be enforced to ensure only approved AI/ML tools can be accessed.	PR.AA-05
11.7.8	Tools **must** execute in sandboxed environments (filesystem/network): default-deny outbound; no shell/exec unless explicitly approved.	PR.PS-01 PR.IR-01

11.7.9	All compiled AIBOM and MLBOM artifacts **must** be integrated directly into the central DWP Asset Registry to ensure continuous, dynamic threat monitoring and unified vulnerability tracking across the enterprise estate.	ID.AM-02 DE.CM-09
11.7.10	To balance secure pipeline hygiene with developer velocity, automated GitLab License Approval Policies shall restrict blocking actions to 'Only New Dependencies'. Where package-level PURL exclusions return an unknown SPDX identifier in feature branches, the pipeline will 'fail open' to allow continued development. However, a 'fail closed' manual override gate **must** be enforced prior to merging into any protected default branch. Approval for this override is strictly a First-Line operational accountability, requiring sign-off from Secure Development in conjunction with the designated Lead Engineer for the respective codebase.	ID.RA-01

11.8 Evaluation, Red-Teaming, Assurance and Incident Response

(Important) this table contains ‘must’ activities.

Reference	Minimum Technical Security Measures	NIST ID
11.8.1	Pre-go-live adversarial testing and red-teaming requirements **must** be applied proportionate to risk. Where red-teaming requirements are required they **must** explicitly evaluate system resilience against prompt injection, guardrail evasion, tool-call manipulation, and the unauthorised extraction of system prompts, steering files, hidden context windows, and Model Context Protocol (MCP) definitions. Detailed penetration-testing methodologies, dynamic test execution playbooks, and adversarial payload libraries **must** conform to SS-027: Security Testing Standard [Ref. K]. Testing of functional model accuracy, statistical drift, and precision/recall metrics **should** be conducted using approved testing frameworks (such as the AI Testing Framework) (see external references) as a complementary assurance activity.	ID.RA-01 ID.IM-02

11.8.2	Performance safety thresholds and automatic fallback/kill-switch criteria **must** be established.	PR.IR-03 PR.IR-04
11.8.3	All model changes (weights, prompts, system messages, guardrails) **must** be approved and recorded (via established change control processes).	ID.RA-07
11.8.4	Red-team results **must** be documented within approved assurance records in line with SS-027 Security Testing Standard [Ref. K].	ID.IM-02
11.8.5	Creation and maintenance of an AI Incident Response IR playbook which **must** contain: • Immediate model rollback/fallback; • tool disablement; • revocation / rotation of credentials; • isolation of poisoned datasets/indices; • purging of caches; • disabling of affected retrieval domains.	ID.AM-02
11.8.6	Incident artefacts (e.g. prompts, tool calls, outputs, indexes) **must** be captured with a chain of custody; the ATRS **must** be updated with incident status/outcome.	RS.AN-03 RS.AN-07
11.8.7	Following an AI/ML security incident or material red-team exercise findings, remediation actions **must** be implemented and recorded. Relevant security testing **must** be repeated to verify that identified weaknesses have been addressed and that updated prompts, system messages, guardrails, credentials, and security controls remain effective.	RS.AN-03

11.8.8	AI/ML services **must** be re-evaluated if increased threat or risk levels are identified in addition to material architectural changes to the model, supplier arrangements, operating environment, or intended use. Continued operation **must** depend on the risk posture remaining acceptable. Where an external provider changes a model, API, safety configuration, or hosting arrangement, the DWP system owner **must** assess whether re-assurance, regression testing, ATRS update, DDA review, risk acceptance or review of the DPIA is required before continued or expanded use.	ID.AM-08
11.8.9	AI systems **must not** progress to production citizen-facing use or consequential citizen-affecting use until conformance with this standard has been evidenced, any exceptions have been formally approved, approval has been granted through appropriate governance boards and assurance requirements proportionate to risk have been completed and evidenced and the following activities have been completed: • Security assurance, • AI-specific evaluation, • Independent IT Health Checks (ITHC) proportionate to risk **must** be conducted in accordance with SS-027: Security Testing Standard [Ref. K], • Monitoring readiness, • Fallback/rollback arrangements, • Incident response procedures, • ATRS requirements, • DPIA, • Support ownership.	ID.RA-01

11.8.10	Monitoring AI / ML services **must** consider the following: • Model performance (for example as defined in service levels), • Safety characteristics, • Behavioural drift over time, • Degradation in accuracy, • Safety, • Grounding, • Policy compliance. When deviations are detected this **must** trigger investigation, reassessment, and corrective action.	DE.CM-09
11.8.11	A review by the Responsible Artificial Intelligence Steering Group (RAISG) **must** be triggered when any of the following conditions arise: • A new model family is introduced. • An AI provider changes. • An agent gains additional privileges. • A new tool connector is added. • A data classification changes. • A RAG corpus scope materially changes. • An AI system becomes citizen facing.	ID.RA-07

11.9 ATRS Registration & Lifecycle Controls

Reference	Minimum Technical Security Measures	NIST ID
11.9.1	AI/ML systems **must** be assessed against the Cabinet Office Algorithmic Transparency Recording Standard (ATRS) criteria prior to production deployment. Where a use case falls within ATRS publication criteria, the system **must** be registered and maintained on the Departmental register, responsibility for which lies with the Senior Responsible Owner (who may be the System Owner or project Sponsor). Where exempt under approved Departmental policy (as administered by the CDO team), the formal exemption rationale **must** be recorded within the System Assurance Record.	ID.AM-02

11.9.2	The ATRS entry for the AI/ML product **must** be maintained accurately throughout the complete product lifecycle, including changes in capability, data, integration pattern, assurance status, and intended use.	ID.AM-02
11.9.3	ATRS entries **must** include as a minimum: • the AI/ML purpose/use-case; • the model family/version; • data categories/classifications processed; • the DPIA reference where applicable; • the security owner; • current assurance status; • supplier details; • data and AI/ML compute hosting location; • evaluation results; • red-team status; • human-in-the-loop design; • monitoring controls; • exception IDs (if applicable). To manage public transparency without exposing operational vulnerability, the entry **must** segment public disclosure into summary information (Tier 1) and protected technical records (Tier 2) [See Scope].	ID.AM-02
11.9.4	Changes to model, data scope, or risk posture **must** trigger ATRS update and (where material) Digital Design Authority review.	ID.AM-02
11.9.5	ATRS compliance or validated approved exemption **must** be a gating criterion for go-live and continued operation. Non-compliance with ATRS requirements requires an approved exception by Digital Design Authority.	ID.AM-02

11.9.6	AI/ML services and supporting artefacts **must** be retired when no longer required. Retirement **must** include, where applicable: • Revocation of user service and API access; • Disabling of integrations and tool connectors; • Retirement or archiving of model artefacts; • Deletion or approved archiving of prompts, • Evaluations; • Embeddings; • Vector indexes; • Caches; • Updates of AI inventory and ATRS records; • Retention of required assurance and audit evidence. Sanitisation and destruction of storage assets **must** comply with SS-036 Secure Sanitisation and Destruction Standard [Ref. L].	ID.AM-02 ID.AM-08 PR.PS-02

11.10 Retrieval-Augmented Generation (RAG) Security

(Important) this table contains ‘must’ activities.

Reference	Minimum Technical Security Measures	NIST ID
11.10.1	Deny-by-default for cross-domain fetch **must** be enforced; allowlist explicit domains.	PR.DS-10
11.10.2	Grounded citation checks **must** be implemented and require linked sources for actions such as but not limited to payments, citizen-facing systems and consequential citizen-affecting outcomes.	PR.PS-01 PR.PS-06
11.10.3	Poisoning indicators **must** be monitored and alerted upon; • sudden topic drift, • toxic markers, • embedded prompt instructions, and • anomalous doc churn.	DE.CM-09

11.10.4	Retrieved content **must** be subjected to groundedness, provenance, and output-binding checks before it is used to support a response, recommendation, or downstream tool execution. Where policy or risk thresholds require authoritative grounding, ungrounded or unverifiable outputs **must** be blocked, flagged, or routed for human review.	PR.PS-01 PR.PS-06
11.10.5	RAG services **must** monitor and alert on; • poisoning and retrieval abuse indicators, including sudden topic drift, • anomalous document churn, • contradictory source behaviour, • hidden or embedded instructions, • unapproved source changes, • abnormal retrieval patterns, • groundedness degradation.	DE.CM-09
11.10.6	Content added to retrieval corpora **must** be ingested only from approved and authenticated sources. Provenance, ownership, classification, handling instructions, dissemination restrictions and applicable retention requirements **must** be recorded and preserved throughout ingestion, indexing, embedding, retrieval and deletion. Content subject to an applicable originator instruction that prohibits or restricts processing by a non-governmental or third-party service **must not** be added to a retrieval corpus, vector store, embedding service or associated AI/ML processing environment hosted by such a service without documented authorisation from the originator or an authority formally empowered to vary the restriction.	ID.RA-07 PR.DS-10
11.10.7	Where incorrect, poisoned, superseded, or unauthorised retrieved content is identified, the affected content **must** be quarantined, removed, or corrected without undue delay. Re-indexing and restoration **must** be subject to approval, evidence of remediation, and verification that the offending content no longer influences retrieval.	PR.PS-02

11.10.8	AI/ML retrieved content **must not** be treated as trusted control input, as it may contain hidden adversarial instructions. AI/ML systems **must** separate retrieved content from system and orchestration instructions and **must not** permit retrieved content alone to authorise privileged behaviour or downstream tool execution.	PR.PS-06
11.10.9	RAG pipelines deployed across multi-tenant or multi-department environments **must** enforce tenant-level namespace or instance isolation within the vector database, except where shared services or shared databases are utilised. Relying solely on application-level metadata filtering for multi-tenant data access control is prohibited.	PR.DS-01 PR.PS-01 PR.PS-06
11.10.10	Vector databases and embedding repositories processing DWP data **must** be isolated within secure network segments in line with SS-006 Security Boundaries Standard [Ref. M]. Access to embedding databases **must** be protected with strong authentication and cryptographic controls in line with SS-007 Use of Cryptography Standard [Ref. C] to prevent embedding inversion attacks, which allow adversaries to reconstruct sensitive source text from vector data.	PR.AA-03 PR.DS-01 PR.PS-01 PR.PS-06

11.11 Secure End-Of-Life and Decommissioning

(Important) this table contains ‘must’ activities.

Reference	Minimum Technical Security Measures	NIST ID
11.11.1	When an AI system, model, or training database is decommissioned, the system owner **must** establish a formal decommissioning plan in coordination with the relevant data owner. The plan **must** ensure that all training datasets, fine-tuning checkpoints, model weights, and embedding directories are securely deleted.	PR.PS-02
11.11.2	All physical storage media and virtual environments that have stored DWP model weights or claimant training data **must** be sanitised or physically destroyed in compliance with DWP SS-036 Secure Sanitisation and Destruction Standard [Ref. L].	ID.AM-08 PR.PS-02 PR.PS-03

12 Appendices

Appendix A Security Outcomes

The minimum security measures defined in this standard contribute to the achievement of security outcomes described in the table below. For consistency, the official NIST Sub-category IDs have been carried through to standards.

Table 1 – List of Security Outcomes Mapping

NIST Ref	Security Outcome (sub-category)	Related Security measure
GV.OC-01	The organisational mission is understood and informs cybersecurity risk management	11.1.14
GV.OC-02	Internal and external stakeholders are understood, and their needs and expectations regarding cybersecurity risk management are understood and considered	11.1.14
GV.OC-03	Legal, regulatory, and contractual requirements regarding cybersecurity - including privacy and civil liberties obligations - are understood and managed	11.1.14
GV.RR-02	Roles, responsibilities, and authorities related to cybersecurity risk management are established, communicated, understood, and enforced	11.1.1, 11.1.16, 11.4.10
GV.PO-01	Policy for managing cybersecurity risks is established based on organisational context, cybersecurity strategy, and priorities and is communicated and enforced	11.1.12
GV.PO-02	Policy for managing cybersecurity risks is reviewed, updated, communicated, and enforced to reflect changes in requirements, threats, technology, and organisational mission	11.1.1

GV.SC-02	Cybersecurity roles and responsibilities for suppliers, customers, and partners are established, communicated, and coordinated internally and externally	11.1.12
GV.SC-05	Requirements to address cybersecurity risks in supply chains are established, prioritised, and integrated into contracts and other types of agreements with suppliers and other relevant third parties	11.5.8
GV.SC-07	The risks posed by a supplier, their products and services, and other third parties are understood, recorded, prioritised, assessed, responded to, and monitored over the course of the relationship	11.3.8, 11.7.3
ID.AM-02	Inventories of software, services, and systems managed by the organisation are maintained	11.1.15, 11.1.20, 11.7.2, 11.7.9, 11.8.5, 11.9.1, 11.9.2, 11.9.3, 11.9.4, 11.9.5, 11.9.6
ID.AM-07	Inventories of data and corresponding metadata for designated data types are maintained	11.1.4, 11.4.2, 11.7.2
ID.AM-08	Systems, hardware, software, services, and data are managed throughout their life cycles	11.1.11, 11.1.15, 11.3.2, 11.3.9, 11.5.7, 11.8.8, 11.9.6, 11.11.2
ID.RA-01	Vulnerabilities in assets are identified, validated, and recorded	11.1.18, 11.5.5, 11.7.4, 11.7.5, 11.7.10, 11.8.1, 11.8.9
ID.RA-04	Potential impacts and likelihoods of threats exploiting vulnerabilities are identified and recorded	11.1.2, 11.1.3
ID.RA-05	Threats, vulnerabilities, likelihoods, and impacts are used to understand inherent risk and inform risk response prioritisation	11.1.3

ID.RA-07	Changes and exceptions are managed, assessed for risk impact, recorded, and tracked	11.1.11, 11.8.3, 11.8.11, 11.10.6
ID.RA-09	The authenticity and integrity of hardware and software are assessed prior to acquisition and use	11.7.4
ID.IM-02	Improvements are identified from security tests and exercises, including those done in coordination with suppliers and relevant third parties	11.8.1, 11.8.4
PR.DS-01	The confidentiality, integrity, and availability of data-at-rest are protected	11.2.4, 11.3.10, 11.3.12, 11.5.1, 11.5.2, 11.5.9, 11.5.10, 11.7.1, 11.10.9, 11.10.10
PR.DS-02	The confidentiality, integrity, and availability of data-in-transit are protected	11.3.3, 11.4.5
PR.DS-10	The confidentiality, integrity, and availability of data-in-use are protected	11.1.9, 11.4.6, 11.10.1, 11.10.6
PR.AT-02	Individuals in specialised roles are provided with awareness and training so that they possess the knowledge and skills to perform relevant tasks with cybersecurity risks in mind	11.1.13, 11.4.10
PR.AA-01	Identities and credentials for authorised users, services, and hardware are managed by the organisation	11.2.6, 11.2.7, 11.2.9, 11.2.10, 11.3.4
PR.AA-02	Identities are proofed and bound to credentials based on the context of interactions	11.2.3, 11.2.6, 11.2.8, 11.3.4
PR.AA-03	Users, services, and hardware are authenticated	11.2.1, 11.2.2, 11.2.5, 11.2.6, 11.5.6, 11.10.10
PR.AA-04	Identity assertions are protected, conveyed, and verified	11.2.10

PR.AA-05	Access permissions, entitlements, and authorisations are defined in a policy, managed, enforced, and reviewed, and incorporate the principles of least privilege and separation of duties	11.1.6, 11.1.10, 11.2.5, 11.2.11, 11.4.1, 11.4.3, 11.5.4, 11.7.7
PR.PS-01	Configuration management practices are established and applied	11.1.7, 11.1.19, 11.3.6, 11.3.9, 11.3.10, 11.5.10, 11.7.8, 11.10.2, 11.10.4, 11.10.9, 11.10.10
PR.PS-02	Software is maintained, replaced, and removed commensurate with risk	11.3.2, 11.9.6, 11.10.7, 11.11.1, 11.11.2
PR.PS-03	Hardware is maintained, replaced, and removed commensurate with risk	11.11.2
PR.PS-04	Log records are generated and made available for continuous monitoring	11.1.5, 11.1.17, 11.4.2, 11.4.4, 11.6.2, 11.6.3, 11.6.4, 11.6.6
PR.PS-05	Installation and execution of unauthorised software are prevented	11.3.7, 11.3.14
PR.PS-06	Secure software development practices are integrated, and their performance is monitored throughout the software development life cycle	11.1.6, 11.3.5, 11.3.6, 11.3.7, 11.3.10, , 11.3.14, 11.3.15, 11.4.7, 11.4.8, 11.4.9, 11.10.2, 11.10.4, 11.10.8, 11.10.9, 11.10.10
PR.IR-01	Networks and environments are protected from unauthorised logical access and usage	11.1.7, 11.1.17, 11.3.1, 11.3.13, 11.7.6, 11.7.8, 11.3.15, 11.3.16
PR.IR-03	Mechanisms are implemented to achieve resilience requirements in normal and adverse situations	11.1.8, 11.3.11, 11.8.2

PR.IR-04	Adequate resource capacity to ensure availability is maintained	11.3.11, 11.8.2
DE.CM-09	Computing hardware and software, runtime environments, and their data are monitored to find potentially adverse events	11.2.12, 11.5.3, 11.6.3, 11.6.4, 11.6.5, 11.7.9, 11.8.10, 11.10.3, 11.10.5
RS.AN-03	Analysis is performed to establish what has taken place during an incident and the root cause of the incident	11.8.6, 11.8.7
RS.AN-07	Incident data and metadata are collected, and their integrity and provenance are preserved	11.8.6

Appendix B Internal References

Below, is a list of internal documents that should be read in conjunction with this standard.

Table 2 – Internal References

Ref	Document	Publicly Available*
A	SS-001 Part 1 Access & Authentication Standard	Yes
B	SS-002 PKI & Key Management Security Standard	Yes
C	SS-007 Use of Cryptography Security Standard	Yes
D	SS-033 Patching & Exposure Management Standard	Yes
E	SS-011 Containerisation Security Standard	Yes
F	SS-003 Software Development Security Standard	Yes
G	DWP Artificial Intelligence Security Policy	Yes
H	SS-020 Data Loss Prevention Security Standard	Yes
I	DWP Data Obfuscation Security Policy	Yes
J	SS-012 Protective Monitoring Security Standard	Yes
K	SS-027 Security Testing Standard	
L	SS-036 Secure Sanitisation and Destruction Standard	Yes
M	SS-006 Security Boundaries Standard	Yes
N	DWP Security Assurance Strategy	No
O	DWP Architecture Blueprint (technology radars)	No
P	SS-001 Part 2 Privileged User Access Standard	Yes
Q	SS-023 Cloud Security Standard	Yes
R	DWP Data Protection Policy	Yes
S	AI Testing and Assurance Engineering Standard	No

Requests to access non-publicly available documents **should be made to an assigned DWP Security Architect or DWP Contracts/Supplier Manager.*

Appendix C External References

The following publications and guidance were considered in the development of this standard and **should** be referred to for further guidance.

Table 4 – External References

External Documents List
NCSC. Guidelines for Secure AI System Development (secure design, development, deployment, and operation).
DSIT. AI Cyber Security Code of Practice and Implementation Guide.
UK Government Security Group. *Artificial Intelligence – UK Government Security resources*.
NIST. AI RMF 1.0 and Generative AI Profile.
NIST IR 8596: Cybersecurity Framework Profile for Artificial Intelligence
NIST SP 800-53 Rev.5 – Security and Privacy Controls.
NIST SP 800-161 Rev.1 – Cybersecurity Supply Chain Risk Management.
NIST SP 800-218 – Secure Software Development Framework (SSDF).
ISO/IEC 23894:2023 – AI Risk Management.
ISO/IEC 42001:2023 – AI Management System (AIMS).
OWASP Top 10 for LLM Applications / OWASP GenAI Security Project.
OWASP: Top 10 Risk & Mitigations for LLMs and Gen AI Apps
AI Playbook for the UK Government
AI Testing Framework
NCSC: Managing the cyber risk of agentic AI
Medicines and Healthcare products Regulatory Authority (MHRA) Software as a Medical Device (SaMD) regulations
ETSI EN 304 223: Security Artificial Intelligence (SAI); Baseline Cyber Security Requirements for AI Models and Systems

Appendix D Abbreviations

Table 4 – Abbreviations

Abbreviation	Definition
AI	Artificial Intelligence
ATRS	Algorithmic Transparency Recording Standard
DPIA	Data Protection Impact Assessment
DSIT	Department for Science, Innovation & Technology
LLM	Large Language Model
ML	Machine Learning
BOM	Bill of Materials
ML-BOM	Machine Learning Bill of Materials
MLOps	Machine Learning Operations
NCSC	National Cyber Security Centre
PDU	Product Delivery Unit
RBAC	Role-Based Access Control
SSDF	Secure Software Development Framework

Appendix E Definition of Terms

Table 5 – Glossary

Term	Definition
Adversarial Attack	Inputs crafted to induce misclassification or unsafe behaviour.
Agentic AI	An AI system or 'agent' capable of planning, making decisions, invoking tools, interacting with systems or executing actions with varying degrees of autonomy in pursuit of an objective.
AI/ML System	Any model, agent, inference service, training pipeline, or decision support tool using AI/ML.
AI Bill of Materials	a standardized, machine-readable inventory of every component that makes up an Artificial Intelligence system.
Algorithmic Transparency Recording Standard	A UK government framework which allows public sector organisations to publish details about how and why they use artificial Intelligence and algorithmic tools which may influence public decision-making or directly interact with citizens. The use of the ATRS is mandatory for central government departments.
Caldicott Guardian	The person responsible for protecting the confidentiality of people's health and care information, ensuring compliance with MHRA and health data governance.
Canary	An early warning indicator designed to detect security breaches, model degradation, or transformative shifts in technology.

Consequential citizen-affecting decisions or outcomes	• Using an AI tool to process personal data • Introducing an AI tool into an existing business process which involves processing personal data • Using the output of an AI tool in a way that will affect people. It is important to note that this applies whether the 'consequence' is a positive or a negative one.
Data Poisoning	Malicious manipulation of training data.
Data Lineage	The process of mapping and tracking a dataset's journey to its final use in training or prompting an AI model.
Excessive Agency	An AI agent or language model which is granted too much autonomy, authority, or system access. Leading to a critical security vulnerability where an AI can autonomously execute damaging or unauthorized actions due to a misinterpreted prompt, hallucination, or manipulation.
Git-Tree	A git-tree is the current snapshot of a project which contains all files and directories which are currently tracked by GIT in addition to modifications, new files and untracked resources.
Groundedness	The practice of anchoring a model's responses to verified, retrievable source material rather than letting it rely solely on its internal training data.
ML-BOM	A structured inventory that documents all the components used to build, train, and deploy a machine learning model.
Model Context Protocol	The standardisation of how AI applications connect to external data sources, services, and tools.
Model Provenance	Documentation of training, datasets, licences, updates, and evaluation results.
Model Drift	Degradation of performance due to changing data patterns.
Nonce	A nonce (a "number used once") is a unique, random string or number generated for a single, specific message. The receiver records the nonces it has already processed; if a message arrives with a previously used nonce, the server rejects it as a duplicate.
Prompt Injection	Inputs that manipulate model/agent behaviour by overriding instructions.

Provenance	The verifiable history, origin, and chain of custody behind AI systems and the content they generate.
Retrieval Corpora	The external databases or document collections used by AI models to fetch relevant information, overcome knowledge cut-offs, and reduce hallucinations in RAG systems.

Appendix F Accessibility artefacts

A variety of accessibility guidance is available from the below URL, that includes:

DWP Accessibility Manual

Guidance and tools for digital accessibility

Understanding accessibility requirements for public sector bodies

Appendix G AI-Specific Threat Examples (for Threat Modelling)

Threat	Description	Mitigations
Data Poisoning	Poisoned samples cause targeted misclassification.	Dataset curation, Robust training, Outlier detection.
Prompt Injection & Tool Misuse	Malicious content overrides guardrails, induces sensitive data exfiltration or unsafe tool calls.	Content isolation, Allow-listed tools, Output-binding checks, Human-in-the-loop.
Model Extraction/Theft	Query-based replication or weight exfiltration	Rate-limiting, Watermarking, IP protection, Hashing and Cryptographic Signing.
Adversarial Perturbations	Imperceptible changes mislead models (vision/NLP)	Adversarial testing, Robust model architectures.
Model Inversion & Memorisation	Infers training data from outputs.	Privacy-enhancing techniques, Output filters, Audit sampling.
Supply-Chain Compromise	Tainted and malicious models and/or libraries.	Provenance checks, SBOM / MBOM, 3rd-party assurance.
Hallucination-Induced Decision Risk	Inaccurate outputs drive operational error (possibly impacting UK Citizens).	Confidence signalling, Claims grounding, Human approval for high-risk actions.
AI Uplifted Cyber Kill-Chain	Cyber kill-chain activities which have enhanced the ability for threat actors to perform elements of the cyber kill chain such as automated reconnaissance, vulnerability research and exploit development, credential harvesting, privilege escalation, discovery, lateral movement, command-and-control and exfiltration.	Threat modelling, Continous exposure management, Least privilege agent, Workload identified, MCP Server and tool allowlisting, Down-scoped delegated tokens; Restricted network and file systems access; Isolated or sandboxed executions; Behavioural monitoring; Rapid credential and tool revocation; security testing.