NHS England

Cyber Security for AI in the NHS

Contents

1. Purpose and Scope

The (national) NHS Artificial Intelligence (AI) Team have developed this guidance in conjunction with the (national) Joint Cyber Unit (JCU). It draws on publicly available AI and cyber security guidance, including Department of Science, Innovation and Technology (DSIT) and European Telecommunications Standards Institute (ETSI) publications.

The purpose is to provide frontline organisations with practical support, helping them to identify and manage cyber risks that may arise when using AI systems. It focuses on key areas such as secure-by-design, lifecycle controls, data governance, monitoring, and human oversight.

Appreciating the overwhelming amount of cyber and AI guidance publicly available, the authors of this guidance have made a conscious effort to ensure it is as helpful and reader friendly as possible. It does not map or signpost an infinite loop of existing standards and frameworks, nor is it tailored to specific AI use cases. Instead, it aims to help organisations think about the most likely cyber risks that are introduced when adopting AI; what the potential impact could be; and what steps can be taken to reduce the chance of it happening. It is worth noting that AI-specific cyber failures can manifest as clinical safety risks. Organisations should ensure that AI cyber risks which could plausibly affect clinical decisions, prioritisation, or patient communication are jointly reviewed through both cyber and clinical safety governance to ensure they are not seen as an ‘IT problem’.

The risks and controls described in this guidance span the full lifecycle of AI use, from design and development, through deployment and maintenance, to end of life. Where technical terms are used (for example, RAG), links are provided to authoritative external guidance listed in the informative references and appendix (rather than redefining terms within this document).

Please note that AI-enabled cyberattacks and defensive use of AI in cyber are out-of-scope of this guidance. If you have any questions, please contact england.responsible.ai@nhs.net

2. Intended Audience

This guidance is for frontline NHS organisations adopting or operating AI systems. It is most relevant to:

- Cyber and digital teams responsible for the secure deployment, operation, monitoring, and incident response of AI systems.

- Clinical safety and clinical leadership roles (including CCIOs and Clinical Safety Officers), where AI cyber failures could affect clinical decisions, prioritisation, or patient communication.

- Information governance and data protection teams, where AI cyber risks may impact confidentiality, integrity, transparency, and auditability of data.

- Senior responsible owners, executives, and boards accountable for AI adoption, oversight, risk acceptance, and assurance.

- Procurement and supplier management teams involved in sourcing AI systems or third-party models and APIs.

3. Using this Guidance

This guidance is not a checklist. Controls should be applied proportionately, based on the organisation's role (for example, developer, deployer/system operator, procurer, data owner/custodian), the criticality of the use case, and the potential impact on patients, staff, and services.

4. Alignment with Existing Requirements

This guidance is AI-specific cyber guidance intended to help NHS organisations understand and manage the additional cyber risks introduced when deploying and operating AI systems. It sits alongside, and does not replace, existing requirements. The AI and Digital Regulations Service (**AIDRS**) cyber guidance sets out statutory cyber resilience duties under the Network and Information Systems (**NIS**) Regulations that apply to all digital services, with compliance evidenced through the Data Security and Protection Toolkit (**DSPT**), which provides organisational level assurance that appropriate cyber and IG (information governance) controls are in place. The Digital Technology Assessment Criteria (**DTAC**) acts as a procurement and due diligence gateway, ensuring digital technologies meet baseline standards for clinical safety, data protection and technical security before adoption. Where an AI system meets the definition of a medical device, Medicines and Healthcare products Regulatory Agency (**MHRA**) requirements apply, regulating safety, performance and lifecycle oversight of the product itself. Where AI systems are used in clinical contexts, organisations should also consider their obligations under the NHS clinical safety standards **DCB0160** (for deployment and use) and, where applicable, **DCB0129** (for manufacture), which address clinical risk management rather than cyber.

The table below summarises how this guidance fits alongside existing requirements relevant to the adoption of AI.

Requirement	Purpose	Focus
AIDRS Guidance	Helpful guidance that sets out statutory cyber resilience duties	Legal cyber obligations (e.g. NIS), incident reporting, organisational resilience
DSPT	Provides organisational assurance for handling NHS patient data	Evidence that a minimum set of cyber and IG controls are in place
DTAC	Acts as a procurement gateway	Baseline checks for clinical safety, data protection and technical security when procuring digital technologies
DCB0129	Assures clinical safety during manufacture	Clinical risk management for the development of health IT systems
DCB0160	Assures clinical safety during deployment and use	Clinical risk management of health IT systems in live operational and care settings
MHRA requirements	Regulate medical devices, including AI where applicable	Product safety, performance, risk management and lifecycle oversight
This AI Cyber Guidance	**Addresses AI-specific cyber risks within the NHS**	**AI-specific threats, lifecycle controls, monitoring and assurance**

5. Potential AI Cyber Risks and How to Manage Them

Organisations should be mindful that deploying AI systems where they are not clearly required can unnecessarily increase system complexity, data exposure and attack surface, introducing cyber risk without corresponding benefit. This section sets out the potential cyber

risks that may be introduced when deploying and operating AI systems. For each potential risk you will find four useful pieces of information:

- Potential Risk – The cyber risk that you may unintentionally expose your organisation to when deploying an AI system.
- Potential Impact – Helps you to understand what might happen if the risk were to become a reality.
- Controls to Consider – What actions you can take to try and prevent the risk from occurring.
- Informative References – Further guidance that can help you implement a control.

Roles and Responsibilities:

This guidance uses common AI role terminology to describe where cyber risks may arise across the AI lifecycle and where controls can be applied. These terms are used descriptively to support understanding and do not introduce new roles or responsibilities. An organisation may perform more than one role for a given AI system.

Developer refers to an organisation or team that designs, trains, fine-tunes, or materially modifies an AI model or system. This may include suppliers, third-party service providers, or internal development teams.

Deployer/system operator refers to an organisation that configures, integrates, deploys, operates, or makes an AI system available for use in practice. These terms are used interchangeably in this guidance.

Procurer refers to an organisation that sources or commissions an AI system, model, or AI-enabled service from a third-party.

Data owner/custodian refers to an organisation responsible for the governance, protection, and appropriate use of data used by or accessed by an AI system, regardless of whether it developed or operates the model itself.

Controls in this guidance should be applied proportionately, based on the role or roles performed and the organisation's ability to influence the design, configuration, data, and operation of the AI system.

5.1 Governance & Accountability

Potential Risk: AI systems are deployed or operated without clearly defined cyber accountability, decision rights, or escalation authority, resulting in unclear ownership of AI-specific cyber risks across the system lifecycle.

Potential Impact: Inability to prevent, detect or respond effectively to AI-specific cyber incidents; delayed suspension of unsafe or compromised AI systems; weak assurance evidence; increased exposure to data compromise, service disruption, or regulatory challenge following a cyber incident.

Controls to Consider:

- Assign clear accountability for AI systems, including ownership for security configuration, monitoring, incident response and suspension decisions, proportionate to the organisation’s role (e.g. developer, deployer/system operator, procurer, data owner/custodian).
- Define and document cyber related decision rights for AI systems, including authority to approve deployment, constrain functionality, suspend use, or decommission following cyber risk escalation.
- Ensure AI cyber risks are explicitly considered within existing cyber governance structures (e.g. risk management, incident management, assurance and audit), rather than managed in parallel or informally.
- Ensure personnel with responsibilities for AI security (e.g. threat modelling, configuration, monitoring, incident response) receive proportionate training and maintain competence appropriate to their role.

Informative References:

- ETSI TR 104 128, Section 6.1 — 'Principle 1: Raise awareness of AI security threats and risks.'
- ETSI TR 104 128, Section 6.2 — 'Principle 2: Design the AI system for security as well as functionality and performance.'
- ETSI TR 104 128, Section 6.4 — 'Principle 4: Enable human responsibility for AI systems.'

5.2 AI-Specific Threat Modelling

Potential Risk: AI systems are deployed without structured consideration of AI-specific threats (e.g. data poisoning, prompt injection, model extraction or evasion), leading to incomplete or misaligned security controls.

Potential Impact: Unrecognised attack paths result in exploitation that bypasses traditional cyber controls, causing unsafe outputs, data leakage, service disruption, or loss of trust.

Controls to Consider:

- Perform proportionate, AI-specific threat modelling prior to deployment and when making material changes, covering threats such as data poisoning, indirect prompt injection, model extraction, evasion and misuse.
- Record identified threats, assumptions and mitigations as part of the pre-deployment risk summary and retain for audit and review.
- Revisit threat models following incidents, significant changes, or emergence of new AI relevant threats.

Informative References:

- ETSI TR 104 128, Section 6.3 — 'Principle 3: Evaluate the threats and manage the risks to the AI system.'

5.3 Shadow AI & Missing Inventory

Potential Risk: AI models, services or pilots are introduced or used without being formally recorded in a central AI inventory, assigned a named owner, or supported by sufficient documentation (such as a model card) to enable effective cyber asset management and oversight. This may include the use of non-corporately provided AI tools outside organisational controls.

Potential Impact: Unmanaged data flows, unknown dependencies and access paths, limited monitoring or logging, delayed or ineffective incident response, and reduced ability to suspend, contain or decommission AI systems following cyber compromise or misuse.

Controls to Consider:

- Maintain a central inventory of all AI systems in use (including pilots, locally developed models, externally hosted models and third-party APIs) sufficient to support cyber asset management, monitoring, incident response and assurance.
- Assign a named accountable owner within the organisation for each AI system, responsible for its secure configuration, operation, monitoring, and for initiating suspension or escalation in response to cyber risk.
- Require each AI system to have a model card (or equivalent structured documentation) that records, at a minimum, the system's purpose, intended use, deployment context, key dependencies, data inputs, access constraints, known limitations and security-relevant assumptions.
- Ensure model cards and inventory records are kept up to date across the AI lifecycle and reviewed following material changes to models, data sources, configurations or deployment environments.
- Existing Shadow IT and acceptable-use controls apply equally to AI tools.

Informative References:

- ETSI TR 104 128, Section 6.5 — 'Principle 5: Identify, track, and protect assets.'
- NCSC Shadow IT Guidance. (While not AI-specific, this guidance is relevant to unmanaged systems.)

5.4 Data Provenance & Representativeness Weaknesses

Potential Risk: Insufficient assurance over the origin, integrity or handling of data used to train, fine-tune or operate AI systems enables data poisoning, contamination, or unauthorised modification of inputs.

Potential Impact: Compromised model behaviour, unsafe or misleading outputs, loss of confidence in AI system integrity, and reduced ability to investigate or remediate AI-specific cyber incidents.

Controls to Consider:

- Apply proportionate assurance over data sources used by AI systems, including verification of origin, integrity, access controls and handling across the data supply chain.

- Maintain auditable records of material datasets and data transformations relevant to deployed AI systems to support cyber incident investigation and rollback.
- Protect operational data pipelines and stores against unauthorised access or modification using appropriate access control, monitoring and integrity mechanisms.

Informative References:

- ETSI TR 104 128, Section 6.2 — 'Principle 2: Design the AI system for security as well as functionality and performance.'
- ETSI TR 104 128, Section 6.3 — 'Principle 3: Evaluate the threats and manage the risks to the AI system.'
- ETSI TR 104 128, Section 6.7 — 'Principle 7: Secure the supply chain.'
- ETSI TR 104 128, Section 6.8 — 'Principle 8: Document data, models and prompts.'

5.5 Transparency & Documentation

Potential Risk: AI systems lack sufficient documentation of intended use, system boundaries, configuration and limitations to support secure operation, monitoring, audit and incident response.

Potential Impact: Reduced ability to detect misuse or compromise, ineffective investigation of AI related cyber incidents, inability to demonstrate control or accountability, and delayed containment or recovery following security events. Loss of public trust, complaints and challenges, regulatory enforcement action, and difficulty defending clinical or operational decisions.

Controls to Consider:

- Document the intended purpose, operational boundaries, key assumptions and limitations of AI systems to the extent necessary to support secure deployment, monitoring and incident response.
- Ensure documentation relevant to cyber operation (e.g. system architecture, interfaces, logging, retention, and suspension mechanisms) is accessible to appropriate cyber and assurance functions.
- Maintain documentation as a living artefact, updating it following material security-relevant changes to models, data, configuration or deployment context.

- Provide users with concise guidance on appropriate use, known limitations, and when to escalate or refrain from relying on outputs.

Informative References:

- ETSI TR 104 128, Section 6.8 — 'Principle 8: Document data, models and prompts.'
- ETSI TR 104 128, Section 6.9 — 'Principle 9: Conduct appropriate testing and evaluation.'
- ETSI TR 104 128, Section 6.10 — 'Principle 10: Communication and processes associated with end-users and affected entities.'

5.6 Access Control & Secret Management Failures

Potential Risk: Overprivileged accounts, unmanaged API keys, and poor segregation expose patient data and model artefacts. (This may not fully address scenarios where AI assistants operate using the permissions of the user, enabling broad or automated access to data that the user can access, but would not typically retrieve or process as part of normal activity.)

Potential Impact: Unauthorised access to sensitive data, such as patient data or model weights. Data exfiltration, breach notifications, downtime and costly remediation.

Controls to Consider:

- Enforce least privilege, short lived credentials, secret vaulting, and segregation of duties across pipelines, data stores, inference endpoints and admin tools. Monitor access centrally.
- Apply RBAC for datasets, artefacts and logs, and implement comprehensive logging of access and changes.
- Maintain baseline cyber hygiene (asset inventory, secure configuration, patching, vulnerability management).

Informative References:

- ETSI TR 104 128, Section 6.6 — 'Principle 6: Secure the infrastructure.'

5.7 Supplier & Third-Party Model/API Risk

Potential Risk: External models/APIs are used without assurance of documentation, security controls, usage instructions, and incident processes.

Potential Impact: Misuse due to vague instructions, delayed incident notification, unresolved vulnerabilities and service failure affecting care operations.

Controls to Consider:

- Require technical documentation, clear instructions for use, cyber attestations, logging, and incident reporting commitments. Embed them in contracts.
- Evaluate suppliers against a control checklist (identity, data lineage, config/versioning, monitoring, suspension/reporting).
- Obtain assurance, proportionate to risk, of model and dataset provenance, including declared training sources, version integrity and known limitations, particularly for foundation, fine-tuned or open-source models.

Informative References:

- ETSI TR 104 128, Section 6.7 — 'Principle 7: Secure the supply chain.'

5.8 Model Lifecycle & Configuration Control

Potential Risk: AI models, prompts, datasets or configurations are changed, updated or replaced without adequate lifecycle control, versioning or approval, leading to integrity loss, unpredictable behaviour, or the introduction of new cyber risks in production.

Potential Impact: Undetected regressions, degraded security posture, inconsistent or unsafe outputs, audit gaps, inability to trace root cause, and reduced ability to rollback or contain incidents—potentially impacting patient care and operational services.

Controls to Consider:

- Implement ModelOps/MLOps lifecycle controls for AI systems, including version pinning, environment segregation, configuration baselines and immutable audit logs covering models, prompts, datasets and retrieval sources.

- Require formal change control and approval for material updates to models, data inputs, prompts or configurations prior to deployment, with clear rollback capability and documented decision authority.
- Ensure pre-deployment checks proportionate to cyber risk are completed for changes (e.g. integrity validation, dependency review, access scope review), and that evidence is retained for audit.
- Maintain reproducible deployment artefacts and version history to support incident investigation, rollback and assurance, including alignment between development, training, testing, staging and production environments.
- Apply equivalent security controls to development, training, testing and staging environments as production, including access control, dataset integrity, change logging and segregation of duties, to reduce the risk of poisoning, contamination or unauthorised modification.
- Confirm change approvers and deployers/system operators are trained on AI-specific security risks (e.g. data poisoning, model/config integrity, rollback).

Informative References:

- ETSI TR 104 128, Section 6.1 — 'Principle 1: Raise awareness of AI security threats and risks.'
- ETSI TR 104 128, Section 6.5 — 'Principle 5: Identify, track and protect assets.'
- ETSI TR 104 128, Section 6.6 — ‘Principle 6: Secure the infrastructure.’
- ETSI TR 104 128, Section 6.8 — 'Principle 8: Document data, models and prompts.'
- ETSI TR 104 128, Section 6.12 — 'Principle 12: Monitor the system's behaviour.'

5.9 Prompt/Context Management (LLM/GenAI)

Potential Risk: Unstructured prompts and ingestion of untrusted or externally sourced content (for example emails, documents, web content or multimodal inputs) may introduce hidden or malicious instructions (including indirect prompt injection), leading to unintended disclosure, unsafe outputs, or inappropriate actions by the AI system.

Potential Impact: Sensitive information leakage into outputs, miscommunication to staff/patients, reputational damage and potential legal exposure.

Controls to Consider:

- Define prompt governance: approved system instructions, input filters, and role-based access to context sources; use per subject scoping for retrieval.
- Treat externally sourced or untrusted content as a potential attack vector. Apply proportionate controls such as input filtering or validation, clearly separating trusted system instructions from user or external inputs, and requiring user confirmation before acting on outputs derived from such content.
- Separate trusted instructions from user inputs; log prompts/responses with retention limits and access controls.
- Publish user guidance on safe prompting and sensitive-data handling; highlight disallowed inputs.

Informative References:

- ETSI TR 104 128, Section 6.2 — 'Principle 2: Design the AI System for security as well as functionality and performance.'
- ETSI TR 104 128, Section 6.5 — 'Principle 5: Identify, track and protect assets.'
- ETSI TR 104 128, Section 6.9 — 'Principle 9: Conduct appropriate testing and evaluation.'

5.10 Retrieval Augmented Generation (RAG) Boundary Leakage

Potential Risk: Retrieval from shared stores returns sensitive content beyond minimum necessary or crosses patient/workflow boundaries.

Potential Impact: Cross patient disclosure, IG breach, formal complaints and regulatory scrutiny, erosion of trust and potential financial penalties.

Controls to Consider:

- Enforce row/document level RBAC, query isolation, and per subject scoping; add retrieval audit and query throttling where appropriate.
- Apply ‘minimum necessary’ principles at the retrieval layer by ensuring that only the smallest set of documents, fields or fragments required for the task and subject can be retrieved, and that all retrieval activity is access controlled, logged and auditable.
- Explain retrieval scopes and boundary rules to users.

Informative References:

- ETSI TR 104 128, Section 6.5 — 'Principle 5: Identify, track and protect assets.'
- ETSI TR 104 128, Section 6.9 — 'Principle 9: Conduct appropriate testing and evaluation.'

5.11 Monitoring, Logging & Audit Deficits

Potential Risk: Insufficient monitoring, drift detection and audit logging leads to unsafe or degraded AI behaviour going undetected and weakens organisational accountability.

Potential Impact: Prolonged unsafe operation, delayed remediation, inability to investigate or demonstrate accountability, and heightened regulatory exposure.

Controls to Consider:

- Performance metrics, data/model drift and out-of-distribution indicators.
- Define suspension criteria and escalation paths.
- Retain system logs for an appropriate, risk-based period sufficient to support incident investigation, assurance and oversight. Where logs include prompts, responses or contextual data, organisations should treat these as potentially sensitive data stores and apply appropriate controls for access, retention and governance.

Informative References:

- ETSI TR 104 128, Section 6.11 — 'Principle 11: Maintain regular security updates, patches and mitigations.'
- ETSI TR 104 128, Section 6.12 — 'Principle 12: Monitor the system's behaviour.'

5.12 Assurance, Evidence & Auditability

Potential Risk: AI controls are defined but not supported by retained evidence, limiting the organisation’s ability to demonstrate effective oversight, compliance or safe operation.

Potential Impact: Difficulty responding to incidents, regulatory scrutiny or legal challenge; reduced confidence in AI governance and decision making.

Controls to Consider:

- Retain proportionate assurance artefacts demonstrating that AI cyber controls are defined and operating, such as inventories, approvals, evaluations, logs and monitoring outputs.
- Ensure evidence is accessible to appropriate assurance functions (e.g. cyber, IG, clinical safety, internal audit) and retained in line with organisational policy and risk.
- Periodically review assurance artefacts to confirm continued effectiveness of controls.

Informative References:

- ETSI TR 104 128, Section 6.8 — 'Principle 8: Document data, models and prompts.'
- ETSI TR 104 128, Section 6.10 — 'Principle 10: Communication and processes associated with end-users and affected entities.'

5.13 Human Oversight & Operational Boundary Deficits

Potential Risk: AI systems are deployed without effective mechanisms for human oversight, intervention or suspension when cyber compromise, misuse or unsafe behaviour is suspected.

Potential Impact: Prolonged operation of compromised or misbehaving AI systems, increased impact of AI-enabled cyber incidents, and reduced ability to contain or recover from security events. Unsafe or inappropriate decisions, missed escalation or intervention, patient harm, lack of accountability for outcomes, and increased legal, regulatory and reputational exposure.

Controls to Consider:

- Ensure AI systems are designed and deployed with technical and procedural controls that allow authorised personnel to intervene, constrain functionality or suspend use in response to a cyber incident.
- Define clear operational boundaries for AI systems that support detection of anomalous or unauthorised use and trigger escalation or suspension where required.
- Integrate human oversight mechanisms into cyber incident management and operational resilience processes, rather than relying on informal or ad-hoc intervention.

- Communicate operational boundaries and red-flag conditions to users so they know when to intervene or suspend.

Informative References:

- ETSI TR 104 128, Section 6.4 — 'Principle 4: Enable human responsibility for AI systems.'
- ETSI TR 104 128, Section 6.9 — 'Principle 9: Conduct appropriate testing and evaluation.'

5.14 Model Robustness & Adversarial Resilience Weaknesses

Potential Risk: Models are not assessed for robustness against adversarial, malformed or edge case inputs, resulting in brittle behaviour when exposed to misuse or hostile interaction.

Potential Impact: Unexpected or unsafe outputs, erosion of clinical confidence, automation bias, and increased risk of patient harm or operational disruption.

Controls to Consider:

- Evaluate model behaviour under realistic misuse, edge case and adversarial input scenarios proportionate to risk and intended use.
- Document known failure modes, confidence limitations and conditions under which outputs should not be relied upon.
- Where appropriate, implement input validation, output constraints, confidence signalling or fallback mechanisms to reduce impact of adversarial behaviour.

Informative References:

- ETSI TR 104 128, Section 6.2 — 'Principle 2: Design the AI system for security as well as functionality and performance.'

5.15 Incident Readiness & Suspension Criteria Missing

Potential Risk: No defined criteria or process to suspend AI use and report serious incidents when risks arise.

Potential Impact: Prolonged exposure to risk, wider impact across services, delayed regulatory reporting and larger recovery effort.

Controls to Consider:

- AI-related cyber incidents should be managed through existing organisational cyber incident response and operational resilience arrangements, with additional AI-specific triggers, suspension criteria, and considerations applied where relevant.
- In higher risk contexts, suspend use if risks to health/safety or rights emerge and where necessary report cyber incidents without undue delay (as per DSPT requirements).

Informative References:

- ETSI TR 104 128, Section 6.2 — 'Principle 2: Design the AI system for security as well as functionality and performance.'
- ETSI TR 104 128, Section 6.6 — 'Principle 6: Secure the infrastructure.'
- ETSI TR 104 128, Section 6.10 — 'Principle 10: Communication and processes associated with end-users and affected entities.'

5.16 End of Life (EoL) & Decommissioning Weaknesses

Potential Risk: AI systems are decommissioned or retired without retaining sufficient artefacts, documentation, logs, or contextual information to support later investigation, audit, or challenge. This can result in an inability to explain, justify, or evidence past AI-assisted decisions once the model, configuration, or supporting infrastructure is no longer available.

Potential Impact: Post-retirement breaches via residual credentials/artefacts, reputational damage and unnecessary regulatory exposure.

Controls to Consider:

- Implement retirement runbooks to revoke credentials, validate data deletion (including embeddings, vector stores, caches and backups), archive relevant artefacts and logs, and update inventories when AI systems are decommissioned.
- Retain sufficient documentation, configuration records, logs, and assurance artefacts after decommissioning to support audit, investigation, or challenge relating to past AI-assisted decisions, proportionate to the risk and context of use.

- Ensure that key artefacts needed to understand how the AI system was designed, configured, and operated remain accessible to appropriate assurance functions following retirement, while clearly distinguishing evidence retention from continued system operation.

Informative References:

- ETSI TR 104 128, Section 6.13 — 'Principle 13: Ensure proper data and model disposal.'

Appendix

Authorship, Feedback and Acknowledgements

NHS England welcomes feedback on this guidance. Please contact england.responsible.ai@nhs.net.

This guidance was authored by NHS England, led by the national NHS AI Team in collaboration with the JCU, and reviewed in accordance with NHS England processes. Microsoft Copilot was used by the authors to support the collation and summarisation of publicly available source material listed in this appendix. All content was curated, edited and approved by human authors. All images in this publication are provided under licence.

Other Helpful AI Guidance from NHS England

NHS England has produced the following non-cyber specific AI guidance, which serves as a useful companion to this document. As with the requirements set out in Section 2, this guidance is intended to complement, not duplicate, existing AI guidance.

Ambient scribing guidance for NHS executives/CIOs/CCIOs that emphasises structured documentation, human validation, governance, safety and Information Governance (IG) considerations. See: NHS England – Ambient scribing guidance

And the IG guidance on AI that reinforces lawful data use, transparency, and that clinicians—not AI—make decisions. See: NHS Transformation Directorate – AI IG guidance

UK Government Guidance

1. GDS – AI Playbook (security/governance)
   https://www.gov.uk/government/publications/ai-playbook-for-the-uk-government/

2. DSIT – Code of Practice for the Cyber Security of AI
   https://www.gov.uk/government/publications/ai-cyber-security-code-of-practice/code-of-practice-for-the-cyber-security-of-ai/

3. NCSC – Principles for the Security of Machine Learning
   https://www.ncsc.gov.uk/collection/machine-learning-principles/

4. NCSC – Guidelines for Secure AI System Development
   https://www.ncsc.gov.uk/files/Guidelines-for-secure-AI-system-development.pdf

5. NCSC – Shadow IT
   https://www.ncsc.gov.uk/guidance/shadow-it

European/Regulatory Context

5. EU AI Act – Article 26 (Deployers of high-risk AI systems)
   https://artificialintelligenceact.eu/article/26/

Formal Standards & Technical Specifications

6. ETSI EN 304 223 – Securing AI: Baseline Cyber Security Requirements
   https://www.etsi.org/deliver/etsi_en/304200_304299/304223/02.01.01_60/en_304223v020101p.pdf

7. ETSI TR 104 128 – Securing AI: Implementation Guidance
   https://www.etsi.org/deliver/etsi_tr/104100_104199/104128/01.01.01_60/tr_104128v010101p.pdf

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

9. NIST – AML Taxonomy & Terminology (AI 100-2)
   https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-2e2025.pdf

Threat, Risk & Assurance Knowledge Bases

10. ENISA – Artificial Intelligence Cybersecurity Challenges
    https://www.enisa.europa.eu/publications/artificial-intelligence-cybersecurity-challenges/

11. MITRE ATLAS – Adversarial Threat Knowledge Base
    https://atlas.mitre.org/

12. MIT – AI Risk Repository
    https://airisk.mit.edu/

Industry & Practitioner Frameworks

13. Cloud Security Alliance – AI Controls Matrix (AICM)
    https://cloudsecurityalliance.org/artifacts/ai-controls-matrix/

14. SANS – Critical AI Security Guidelines
    https://www.sans.org/mlp/critical-ai-security-guidelines/

15. OWASP – GenAI Security Project (LLM Top 10)
    https://owasp.org/www-project-top-10-for-large-language-model-applications/

Baseline Cyber Hygiene/Supporting Material

16. CIS – Controls (baseline cyber hygiene)
    https://www.cisecurity.org/controls/

17. Google – Secure AI Framework (SAIF)
    https://blog.google/technology/safety-security/introducing-googles-secure-ai-framework/

18. IBM – Framework for Securing Generative AI
    https://www.ibm.com/products/tutorials/ibm-framework-for-securing-generative-ai/