WP-035
June 2026
12 sections

UK PUBLIC SECTOR AI GOVERNANCE & SOVEREIGN ACCOUNTABILITY

UK Sovereign AI Governance: Cryptographic Proof for Public Sector Accountability

The UK has committed £1.1 billion to sovereign AI infrastructure. 38 of 50 AI Action Plan commitments are delivered. The missing piece is runtime accountability: verifiable proof that public sector AI systems did what they were authorised to do, at the moment they did it.

AffixIO Research | June 2026 | Download PDF

EXECUTIVE SUMMARY

The United Kingdom is building the infrastructure for sovereign artificial intelligence at scale. The AI Opportunities Action Plan, the £500 million Sovereign AI Fund, and the £750 million national AI supercomputer represent a generational commitment to domestic AI capability. The CDDO's Generative AI Framework, the algorithmic transparency standard, the AI Security Institute's evaluation programme, and the NCSC's zero trust principles establish the governance architecture within which that capability must operate. What remains unresolved is the most operationally demanding requirement: runtime accountability.

How does a government department prove, to an auditor, a select committee, or a court, that a specific AI-assisted decision was made by an authorised system, operating within sanctioned parameters, on legitimate data, at the moment in question, not three months later when someone has reconstructed what probably happened from mutable log files? This paper examines that gap, traces it through the NHS, HMRC, and defence contexts, and presents the cryptographic architecture that closes it. AffixIO provides tamper-evident, Merkle-anchored, ML-DSA-65-signed AI interaction records that can demonstrate continuous compliance from evaluation through live deployment, compatible with UK data residency requirements, the Data Protection Act 2018, and the Government Security Classifications framework.

CONTENTS

▶ AffixIO in the UK Sovereign AI Stack
1 The UK's Sovereign AI Moment
2 What Sovereign Actually Means for Governance
3 The CDDO Framework: Ten Principles in Practice
4 Algorithmic Transparency: From Declaration to Proof
5 The AISI Gap: Evaluation Is Not Ongoing Compliance
6 Zero Trust AI: The NCSC Architecture Applied
7 Sector Deep-Dive: NHS, HMRC, and Defence
8 The National Data Library and Sovereign Data Proofs
9 Cryptographic Audit Infrastructure for UK Government
10 Data Residency and the Government Security Classification
11 From Pilot to Sovereign Standard
12 Conclusion: The Audit Trail as National Infrastructure

AFFIXIO : PROVIDER PROFILE

AffixIO in the UK Sovereign AI Stack

The UK sovereign AI programme needs infrastructure at multiple layers: compute to run AI models, data infrastructure to feed them, model development to advance capability, and governance infrastructure to make every AI-assisted decision verifiable and accountable. AffixIO operates exclusively at the governance and accountability layer. We do not train models, operate compute, or hold citizen data. We provide the cryptographic proof infrastructure that makes the rest of the stack accountable: the mechanism by which any AI interaction anywhere in the UK sovereign AI ecosystem can be shown, with mathematical certainty, to have been authorised, attributable, and within sanctioned parameters at the exact moment it occurred.

Where AffixIO Sits in the Stack

STACK LAYER	WHAT IT DOES	PROVIDER TYPE
Compute & infrastructure	GPU clusters, national AI supercomputer, sovereign cloud; runs AI training and inference at scale	DSIT / national supercomputer; hyperscaler UK regions; G-Cloud
AI models & systems	Frontier and domain-specific models; clinical AI; fraud detection; decision-support systems	UK AI companies (Sovereign AI Fund portfolio); global model providers under UK terms
Data & identity	National Data Library; NHS health data; HMRC records; GOV.UK Wallet digital credentials	DSIT; NHS; HMRC; Government Digital Service
Evaluation & safety	Pre-deployment frontier model evaluation; dangerous-capability testing; safety thresholds	AI Security Institute (AISI); NCSC; CDDO

		assurance testing
Governance & proof *← AffixIO*	Runtime cryptographic proof that every AI interaction was authorised, attributable, within policy, and tamper-evidently recorded, compatible with GSC, UK GDPR, DPA 2018, NCSC standards, and CDDO accountability requirements	**AffixIO : domestic UK provider**
Regulatory & legal	ICO enforcement; judicial review; Parliamentary scrutiny; NAO inspection; CQC; FCA; MHRA	Independent regulators and courts; the audit trail AffixIO produces is the evidence layer these bodies require

What AffixIO Provides

Runtime Interaction Records

At the moment of every AI interaction, AffixIO captures a cryptographic record: model version, user identity, data classification, active policy state, and a hash-commitment to the content. No personal data is stored. The record is 200-400 bytes and tamper-evident from the

Merkle-Anchored Audit Batches

Interaction records are assembled into Merkle hash trees. One ML-DSA-65 (NIST FIPS 204) signature covers a million records. Individual records are verifiable via O(log n) inclusion proofs (20 hashes at 1M records per batch). Storage at NHS 8-year retention scale: approximately 18 MB per year in roots and anchors.

Post-Quantum Durability

All batch roots are signed with ML-DSA-65, the NIST-standardised post-quantum signature algorithm. Audit records generated today remain tamper-evident against quantum adversaries for the full regulatory retention period: 8 years for NHS, 12 years for HMRC complex compliance, 30+ years for MOD.

moment of creation.

Zero-Knowledge Authorisation Proofs

A ZK proof demonstrates that an AI interaction used an authorised tool, accessed authorised data, and operated under active controls, without revealing the content of the interaction. Auditors and regulators verify compliance without accessing the AI interaction itself. Satisfies GDPR data minimisation simultaneously with CDDO accountability requirements.

GSC-Aware Deployment

AffixIO deploys across all four Government Security Classification tiers. OFFICIAL and OFFICIAL-SENSITIVE: public cloud or G-Cloud anchoring. SECRET: on-premises or PSNA, HSM key custody. TOP SECRET: air-gapped deployment, closed-network anchoring. The cryptographic protocol is identical across tiers; only the deployment environment changes.

UK Data Residency by Design

AffixIO's audit infrastructure is deployed and operated within UK jurisdiction. Audit records do not transit foreign networks. Signing keys are held in UK-based HSMs. Records are anchored to UK-operated timestamping services or GOV.UK Trust Framework registries. No dependency on foreign vendor co-operation for record retrieval or verification.

AffixIO Capability Map Against UK Frameworks

UK REQUIREMENT	SPECIFIC OBLIGATION	HOW AFFIXIO ADDRESSES IT
CDDO Generative AI Framework, Accountability principle	AI-assisted decisions must be auditable; AI system version must be attributable to each decision	Runtime interaction records bind model version, user identity, and decision timestamp; tamper-evident from point of creation
CDDO Generative AI Framework, Transparency principle	AI systems must be explainable; audit trail must not create secondary data liability	ZK proofs provide evidential compliance without storing interaction content; GDPR data minimisation satisfied simultaneously
Algorithmic transparency standard	AI systems used in public decisions must be declared; runtime behaviour must match declared behaviour	Algorithmic integrity attestation: continuous evidence that the live system matches the transparency record, at model-version granularity
AISI evaluation programme	Frontier models evaluated pre-deployment; no mechanism for continuous post-deployment compliance	Runtime audit trail bridges the AISI gap: ongoing evidence that the live deployed model version is consistent with its evaluated baseline
NCSC Zero Trust, monitoring and instrumentation principle	Monitoring must generate tamper-evident evidence, not merely queryable event streams	Merkle-anchored records are structurally tamper-evident; any post-hoc alteration changes the hash, invalidating the inclusion proof and the signed root
NHS DSPT, immutable audit log requirement	All access to patient data must be immutably logged; logs must be retained for	Merkle-anchored records are cryptographically immutable; selective disclosure via inclusion proofs supports Subject

UK REQUIREMENT	SPECIFIC OBLIGATION	HOW AFFIXIO ADDRESSES IT
	applicable clinical record period	Access Requests without exposing other patients' records
HMRC, judicial review readiness	Algorithmic decisions affecting taxpayers must be reconstructable and challengeable	Per-interaction records with model version, data classification, and timestamp provide the reconstruction trail courts and the Taxpayer Charter require
MOD / GSC SECRET+, post-quantum durability	Audit records must remain tamper-evident against quantum adversaries over 30+ year retention periods	ML-DSA-65 (NIST FIPS 204) signing; on-premises HSM key custody at SECRET; air-gapped deployment at TOP SECRET
UK GDPR Article 15, Subject Access Requests	Individuals affected by AI-assisted decisions must be able to request information about automated processing	Merkle inclusion proofs support selective disclosure: one individual's interaction record can be produced without exposing other records in the same batch
DPA 2018 / UK GDPR Article 5, Data minimisation	AI audit records must not store more personal data than necessary	Records store hash-commitments to interaction content, not content itself; no personal data in the audit trail; content remains in the application layer

What AffixIO Does Not Do

AffixIO does not train AI models, operate compute infrastructure, store citizen or patient data, provide AI model evaluation services, or act as a data processor for government datasets. We are the proof layer, the narrow, technically specialised component of the

sovereign AI stack that makes AI interactions cryptographically accountable. Deploying AffixIO does not replace an organisation's AI system, data platform, cloud infrastructure, or compliance function. It adds the tamper-evident evidence layer that every other part of the governance architecture depends on to produce verifiable outcomes rather than auditable intentions.

SECTION 01

The UK's Sovereign AI Moment

In January 2025, Prime Minister Keir Starmer published the AI Opportunities Action Plan: 50 recommendations to make the United Kingdom a global leader in artificial intelligence. Sixteen months later, 38 of those 50 commitments are formally delivered, and the government has established a live public progress dashboard at delivery.ai.gov.uk to track every remaining milestone. This represents faster delivery than comparable national technology strategies in living memory.

The financial commitments accompanying the plan are substantial. A £1.1 billion investment package announced in early 2026 combines a £750 million commitment to a new national AI supercomputer, direct venture capital through the £500 million Sovereign AI Fund, and skills investment. The Sovereign AI Fund, launched in April 2026 and chaired by James Wise of Balderton Capital, is empowered to take equity stakes in British AI companies, provide access to sovereign compute, and act as an early customer to de-risk investment for the wider market. DSIT's first open procurement round offered up to £80 million to validate AI capabilities in areas identified as national priorities: scientific discovery, health and social care, national security and defence, cybersecurity, transport, and energy.

£1.1bn

committed to UK sovereign AI infrastructure, including £750M for

38/50

AI Action Plan commitments formally delivered by January 2026,

a national AI supercomputer expected operational by 2030

tracked live at delivery.ai.gov.uk

£500m

Sovereign AI Fund launched April 2026 to take direct stakes in UK AI companies and provide compute and data access

30+

frontier AI systems evaluated by the AI Security Institute (AISI) since November 2023, including pre-deployment partnership with Anthropic and others

This is the infrastructure layer. It is impressive, well-resourced, and moving faster than many expected. What it does not yet fully address is the accountability layer: the mechanism by which any individual AI-assisted decision made within this infrastructure can be traced, verified, and demonstrated to have been authorised and within sanctioned parameters at the moment it was made. That accountability layer is where the remaining 12 Action Plan commitments cluster, and it is where this paper focuses.

SECTION 02

What Sovereign Actually Means for Governance

The word "sovereign" in the context of AI has a specific technical meaning that is broader than geography. An AI system is sovereign when:

- **Data residency:** The data it processes does not leave UK jurisdiction without explicit authorisation. For government AI, this is not a preference; it is a mandatory requirement under the Government Security Classifications framework and the Data Protection Act 2018, reinforced by the contractual requirements of the UK GDPR.
- **Compute sovereignty:** The processing infrastructure is owned or controlled by UK entities, such that foreign governments, foreign companies, or foreign legal processes cannot unilaterally access, modify, or disable the infrastructure.

- **Accountability sovereignty:** The governance mechanisms (the audit trail, the accountability framework, the evidence of compliance) are not held by a foreign technology vendor who may be subject to foreign legal compulsion. If HMRC deploys an AI system and the audit records of every decision that system made are held exclusively on foreign servers subject to US CLOUD Act or Chinese national security law, those records are not sovereign.
- **Capability sovereignty:** The UK retains the technical skills, the understanding, and the components to build and maintain AI systems domestically, reducing strategic dependence on a small number of foreign technology providers.

Of these four dimensions, the current investment programme addresses compute sovereignty directly and capability sovereignty through the Sovereign AI Fund's mandate to build British AI companies. Data residency is addressed through procurement requirements and the National Data Library architecture. Accountability sovereignty receives the least attention and is where the most significant gap exists.

> **DEFINITION: ACCOUNTABILITY SOVEREIGNTY**
>
> An AI governance architecture is accountable-sovereign when the evidence of AI system compliance (the audit records, the authorisation chain, the interaction logs) is held in a form that: (a) is cryptographically tamper-evident and does not depend on trusting any specific custodian; (b) can be verified by UK regulatory bodies without requiring co-operation from foreign technology vendors; and (c) remains verifiable over the long term, including against future cryptanalytic advances. Mutable log databases held on foreign cloud infrastructure do not meet this definition.

Accountability sovereignty matters most when things go wrong. An AI-assisted benefits decision that causes material harm to a claimant. An AI fraud detection flag that incorrectly blocks a small business from accessing HMRC services. An AI system deployed in a clinical pathway that contributes to a missed diagnosis. In each of these cases, the question a Parliamentary select committee, the ICO, or a judicial review will ask is: what did the AI

system do, when did it do it, was it authorised to do it, and what were the active safeguards? Without sovereignty over the audit trail, the answer to those questions may not be obtainable from within the UK's own institutions.

SECTION 03

The CDDO Framework: Ten Principles in Practice

The Central Digital and Data Office (CDDO), now integrated with the Incubator for AI (i.AI) and the Government Digital Service under DSIT leadership, published the Generative AI Framework for HM Government. The framework establishes ten principles for government bodies deploying generative AI. It is the most operationally relevant AI governance document in the UK public sector. Two of its ten principles are technically demanding in ways that existing tooling does not fully satisfy.

The Accountability Principle

The CDDO framework requires that human accountability is maintained for AI-assisted decisions, that accountability is documentable and auditable, and that there is a clear chain between the AI system used and the human who bears responsibility for the outcome. In practice, this requires three things that typical enterprise AI deployments do not provide: (1) a record of which specific AI system version made a recommendation, bound to the decision outcome; (2) evidence that the AI system was operating within its authorised parameters at the moment of the recommendation; and (3) an attribution chain that links the AI recommendation to the named civil servant or decision-maker who acted on it. A SIEM log that records "AI system invoked at timestamp X" satisfies none of these three requirements completely.

The Transparency Principle

The CDDO framework requires AI use in government to be documented and, where appropriate, publicly disclosed under the algorithmic transparency standard. The framework also requires that AI systems used in public-facing

decisions are explainable: that users affected by AI-assisted decisions can, on request, receive a meaningful explanation of the factors that influenced the outcome. Explainability and auditability pull in opposite directions in one specific sense: the data that would make an AI recommendation most fully explainable, the full input and output, is often the most sensitive personal data. A compliance architecture that requires storing full interaction transcripts to satisfy transparency creates a secondary data liability. A cryptographic commitment architecture (recording a hash of the interaction, not the interaction itself, with a ZK proof of authorisation) provides the auditable evidence without the content-storage liability.

What the Framework Tests

The CDDO's AI assurance testing framework, described in its September 2025 blog on testing and assuring AI in government, establishes that assurance is proportionate to risk: low-risk AI systems face lighter touch assurance; high-risk AI systems face structured testing and documentation requirements. The framework aligns with the AI Playbook's guidance and the UK AI Safety Institute's evaluation methodology. The key question it leaves open is not "was this system tested before deployment?" but "is this system behaving consistently with its tested baseline during live operation?" Answering that question requires runtime evidence, not pre-deployment evidence.

SECTION 04

Algorithmic Transparency: From Declaration to Proof

The UK's algorithmic transparency standard, launched by the CDDO and Cabinet Office in 2021 and progressively extended since, requires government and public sector organisations using automated decision-making tools to publish transparency records that describe what the tool does, what data it uses, what its intended purpose is, and how human oversight is maintained. As of 2026, the register of published algorithmic transparency records covers tools used across central government, local authorities, and arm's-length bodies.

The standard is valuable. It addresses the legitimacy question: citizens and civil society organisations can see what AI tools their government uses and for what purpose. It does not address the accuracy question: a transparency record states what an AI system is supposed to do. It does not prove what the system actually did on any specific occasion.

This is not a criticism of the standard. It was designed to address legitimacy; accuracy is a different and technically harder problem. But as the number of AI-assisted decisions in the public sector grows, HMRC processes millions of tax returns, NHS systems support hundreds of thousands of clinical interactions annually, DWP makes benefits eligibility determinations at similar scale, the accuracy question becomes operationally important.

> **The transparency gap in numbers:** The algorithmic transparency register documents the design and intent of AI systems. It cannot tell Parliament whether the DWP's UC eligibility model made decisions consistent with its stated parameters in the six months following its last documented update. Only a runtime audit trail can answer that question.

Closing the gap between algorithmic transparency declarations and provable runtime compliance requires what might be called algorithmic integrity attestation: a continuous, tamper-evident record that the AI system in live use matches the system described in the transparency record, that it is processing the data it is authorised to process, and that its output recommendations are being reviewed and overridden in the ways the transparency record indicates. This is distinct from model explainability (explaining individual predictions) and from bias auditing (statistical analysis of outputs across demographic groups). It is the evidence layer that links the public commitment in the transparency record to the actual behaviour of the deployed system.

SECTION 05

The AISI Gap: Evaluation Is Not Ongoing Compliance

The AI Security Institute (AISI), established following the AI Safety Summit at Bletchley Park in November 2023, conducts pre-deployment evaluations of frontier AI models for dangerous capabilities: biosecurity risks, offensive cybersecurity capability, autonomous behaviour, and novel catastrophic risks. As of mid-2026, AISI has evaluated over 30 frontier AI systems, including pre-deployment evaluations conducted in partnership with Anthropic, Google DeepMind, Meta, and the US AI Safety Institute.

This is important and world-leading work. The UK is one of very few countries that has built an independent national capability to evaluate frontier AI before it is deployed in sensitive contexts. The limitation is inherent to the evaluation model: evaluation happens once, before deployment. The evaluated system is then deployed, updated, fine-tuned, and integrated with new data sources and retrieval systems, none of which are separately evaluated.

Consider a concrete scenario. A frontier AI model is evaluated by AISI in March 2026 and found to meet the safety thresholds for deployment in a government advisory role. In May 2026, the model provider releases a minor version update; the government's deployment pipeline applies the update automatically. In June 2026, the updated model makes a recommendation that, had the previous version been in use, it would not have made. The question of whether the June 2026 deployment was the evaluated system is unanswerable without cryptographic evidence of the model version at the moment of each interaction.

ASSURANCE MECHANISM	WHAT IT PROVES	WHAT IT CANNOT PROVE	WHEN IT APPLIES
AISI pre-deployment evaluation	System met safety thresholds at evaluation time	System behaves consistently post-deployment; version in live use matches evaluated version	Once, before deployment
Algorithmic transparency record	Organisation's stated intent for AI system	System behaves as stated; data access is as described; human oversight is as claimed	Published at launch; updated periodically
CDDO assurance testing	System met proportionate pre-deployment assurance criteria	Continuous compliance post-deployment; model version integrity	At deployment; proportionate review triggers
Mutable audit log (SIEM)	Events occurred (cannot prove accuracy of records; no tamper evidence)	Records were not modified after the fact; model version at interaction time; authorisation status	Continuous but unverified
Cryptographic runtime audit trail	Specific model version made specific interaction, under specific authorisation state, at specific time; tamper-evident; post-quantum durable	Content of interaction (by design, for privacy)	Continuous; every interaction

The AISI evaluation and the CDDO assurance framework are best understood as the front end of an accountability chain. The cryptographic runtime audit trail is the back end. Without the back end, pre-deployment assurance

proves what a system was supposed to do; it cannot prove what a system actually did. Both are necessary for a complete sovereign AI accountability architecture.

SECTION 06

Zero Trust AI: The NCSC Architecture Applied

The NCSC's zero trust architecture framework for UK government is built on eight design principles. These principles were developed for network and identity security, but they map directly onto the AI governance problem when extended to cover AI tool authorisation and interaction verification.

The core principle of zero trust is: do not grant trust implicitly; grant it explicitly, verify continuously, and maintain evidence of each grant. Applied to AI governance, this means:

- **Know your architecture:** Maintain a complete inventory of every AI system in use across the organisation, not just the ones you approved, but the ones actually in use. This is the shadow AI inventory problem, documented in WP-034.
- **Know your users, services, and devices:** Know which staff members are interacting with which AI systems, using which data, through which authorisation chains.
- **Assess user behaviour:** Monitor AI interaction patterns for anomalies that may indicate misuse, data exfiltration through AI prompt injection, or AI systems being queried outside their authorised scope.
- **Use policies to authorise requests:** AI tool access should be governed by explicit, cryptographically enforced authorisation policies, not by network location or SSO group membership alone.
- **Authenticate everywhere:** Authentication at the point of AI interaction, not just at login. The fact that someone authenticated to the corporate network three hours ago should not, in a zero trust architecture, be sufficient to authorise them to use an AI tool with SECRET-classified data.

- **Know your service health:** Monitor AI systems for drift from their evaluated baseline. Model version changes, retrieval system updates, and configuration changes should all trigger re-authorisation checks.
- **Focus on monitoring and instrumentation:** The monitoring layer for AI systems must generate tamper-evident evidence, not merely event streams that can be queried after the fact from mutable databases.
- **Do not trust the network:** An AI system's compliance cannot be assumed from the fact that it is running on government infrastructure. The system must prove its own compliance at runtime.

Translating these eight principles into an operational AI governance architecture gives a specific technical requirement: every AI interaction must generate a signed, tamper-evident record that links the interaction to the authorised identity, the specific AI system version, the data classification context, the active policy state, and the timestamp. This record must be generated at runtime, not reconstructed from logs. It must be verifiable by any authorised party without trusting the custodian of the records. And it must remain verifiable over the regulatory retention period, even if the cryptographic primitives in use today are broken by future advances. This is precisely the architecture described in Sections 9 and 10.

SECTION 07

Sector Deep-Dive: NHS, HMRC, and Defence

The UK's three largest public sector AI deployments in 2026 share a common accountability requirement but face it in different regulatory contexts. Understanding each context clarifies what an accountable-sovereign AI architecture must deliver in practice.

NHS: Clinical AI and the DTAC

The NHS has committed to deploying AI across clinical pathways: diagnostic imaging, clinical correspondence, triage, patient flow management, and medicines optimisation. The Digital Technology Assessment Criteria (DTAC),

updated in February 2026, governs the procurement and deployment of AI clinical tools. For Software as a Medical Device, the MHRA's regulatory framework applies in parallel.

NHS audit requirements for clinical AI are more demanding than general CDDO requirements. Every AI-assisted clinical recommendation that influences a patient outcome must be attributable: to the AI system version, to the clinical data accessed, to the clinician who acted on it, and to the time of the recommendation. This record must be retained as part of the patient's care record, which NHS trust policy requires to be retained for a minimum of 8 years for adults and until age 25 for records relating to children. The NHS Data Security and Protection Toolkit (DSPT) requires immutable audit logging for all access to patient data, a requirement that extends to AI systems accessing that data. For clinical AI, mutable logs held in a trust's own SIEM system are not sufficient; tamper-evident records are required by the combination of DSPT requirements, GDPR accountability obligations, and CQC inspection standards.

HMRC: Sovereign AI at Tax Scale

HMRC's May 2026 contract with Quantexa (valued at £175 million over ten years) is the largest public sector sovereign AI deployment announced in the UK to date. The contract commits to building a "fully governed" AI and data analytics platform specifically designed for tax compliance, fraud detection, and entity resolution across HMRC's data assets. The Quantexa platform is explicitly characterised as "sovereign": UK data residency, UK-based support, and contractual commitments that the AI system cannot be directed or inspected by foreign governmental authority.

HMRC's accountability requirements are shaped by Commissioners for Revenue and Customs Act 2005 obligations, the Taxpayer Charter, judicial review risk (taxpayers whose affairs are affected by AI-assisted decisions have judicial review rights), and NAO inspection scrutiny. The NAO has previously flagged concerns about HMRC's ability to demonstrate that algorithmic systems used in compliance work are operating as designed. For the Quantexa deployment to satisfy these requirements, HMRC needs a runtime audit architecture that can demonstrate, for any individual case

affected by the AI system, the specific model version, the data accessed, and the active control configuration, not from reconstructed logs, but from tamper-evident records generated at the moment of each interaction.

Defence and National Security

The Ministry of Defence's AI programmes, across logistics, intelligence analysis, autonomous systems, and decision support, operate under the Government Security Classification framework at SECRET and above. The accountability requirements at these classification levels are more demanding than in any other public sector context: auditability requirements extend to post-quantum durable records, meaning that audit evidence generated today must remain tamper-evident even if a quantum computer capable of breaking current cryptographic standards becomes available within the next 10-20 years.

The NCSC's guidance on post-quantum cryptography is explicit: organisations should be migrating to NIST-standardised post-quantum algorithms now, with a target of critical systems being PQC-ready by 2030. For MOD AI audit records, this means records must be signed with ML-DSA-65 (NIST FIPS 204) or an equivalent post-quantum signature scheme, rather than with ECDSA or RSA, which are vulnerable to quantum cryptanalysis. ML-DSA-65 signatures of 3,293 bytes per root (with Merkle tree batching covering millions of records) provide the post-quantum durability that defence audit requirements demand, with manageable storage overhead using the sublinear attestation architecture described in WP-032.

SECTOR	GOVERNING FRAMEWORK	RETENTION REQUIREMENT	KEY ACCOUNTABILITY OBLIGATION	PQC REQUIREME
NHS clinical AI	DTAC, MHRA SaMD, DSPT, CQC	8 years (adult); to age 25 (child)	Attributable AI recommendation per patient interaction; DSPT immutable log requirement	Recommen (NCSC guidance); mandatory records intended to persist to 2030+
HMRC AI	CRCA 2005, Taxpayer Charter, UK GDPR, NAO	6 years (tax records); 12 years (complex compliance)	Demonstrable system-version attribution for algorithmic decisions; judicial review readiness	Recommen mandatory records subject to judicial revi beyond 20
MOD / national security	GSC SECRET/TOP SECRET, MOD JSP 440, NCSC	30+ years (some categories indefinite)	Post-quantum durable tamper evidence; full classification-aware audit; foreign-adversary-resistant records	Mandatory (NCSC CNS 2.0 timeline
DWP benefits AI	Social Security Acts, Human Rights Act, judicial review	6 years minimum	Explainability for affected individuals; Human Rights Act Article 6 fair trial implications	Recommen
Local government AI	Localism Act, UK GDPR, CDDO guidance	Varies (3-10 years)	Algorithmic transparency standard compliance; council duty of care	Optional in near term

SECTION 08

The National Data Library and Sovereign Data Proofs

One of the most significant commitments in the AI Opportunities Action Plan is the National Data Library: a secure, governed single access point to UK public sector data assets, making HMRC records, NHS datasets, and other national data available for AI development and research under appropriate controls. The Health Data Research Service is the most advanced component, with £600 million committed jointly by the government and the Wellcome Trust to create secure access to NHS data across England, Scotland, Wales, and Northern Ireland.

The governance challenge for the National Data Library is acute. The data it holds is among the most sensitive data in the UK: tax records, medical histories, benefit claims, criminal justice records. The value of this data for AI development is directly proportional to its sensitivity. An AI model trained on linked NHS-HMRC data can identify health conditions correlated with income trajectories, or benefit claim patterns correlated with medical histories, with research value that would otherwise require decades of dedicated data collection. The risk of re-identification, of data misuse, of access beyond authorised scope, is proportional to the same sensitivity.

Zero-knowledge proofs provide a technically specific answer to part of this challenge that is distinct from conventional access controls. A ZK proof allows a researcher or an AI model to demonstrate that a computation over sensitive data has a specific property, for example, that the model was trained on a dataset that meets minimum size thresholds for each demographic group (relevant for bias testing), or that an individual's data was processed only in aggregate and was not singled out, without revealing the underlying data. This is the privacy-preserving computation property, sometimes called "data-in-use" privacy, and it is the specific property that makes ZK proofs relevant to the National Data Library architecture beyond what differential privacy or secure multi-party computation alone can achieve.

AffixIO's role in this context is not as a data custodian. The National Data Library requires sovereign data infrastructure at a scale and sensitivity that is properly the function of government-contracted or government-owned

systems. AffixIO's contribution is the proof layer: the cryptographic attestation that any AI computation performed against National Data Library assets was authorised, within scope, and performed by the specific model version that was granted access, without requiring that the data itself be disclosed to the auditor.

SECTION 09

Cryptographic Audit Infrastructure for UK Government

The architecture that satisfies the CDDO accountability principle, the AISI runtime gap, the NHS DSPT immutable log requirement, the HMRC judicial-review standard, and the MOD post-quantum durability requirement is a single unified design, deployed in configurations appropriate to each classification level.

Interaction Records

At the moment of each AI interaction, the instrumentation layer records:

- A cryptographic hash of the interaction content (binding without disclosure)
- The AI system identifier: model name, version, deployment instance
- A reference to the authorisation registry entry for this model version
- The authenticated user identity (or service account, for agent interactions)
- The data classification of the context (OFFICIAL, OFFICIAL-SENSITIVE, SECRET, or TOP SECRET)
- A commitment to the active policy and safeguard configuration
- A timestamp from a trusted time source (aligned with HMRC and NCSC timestamping requirements)

This record is 200–400 bytes per interaction. It contains no personal data and no interaction content. The personal data and content remain in the application layer; the governance evidence enters the audit trail. This separation satisfies the GDPR data minimisation principle (Article 5(1)(c) UK GDPR) while simultaneously satisfying the accountability and audit requirements of the CDDO framework.

Merkle Tree Anchoring

Interaction records are batched into Merkle hash trees at configurable intervals. The root of each tree is signed with ML-DSA-65 (NIST FIPS 204) and published to an external anchoring service, either a public blockchain, an RFC 3161 qualified timestamping authority, or a government-operated anchoring registry. The batch root can be published to the GOV.UK Trust Framework's trust registry, providing a reference point that is independently verifiable without requiring access to the full interaction record set.

Individual interaction records are verified using Merkle inclusion proofs. At 1 million interactions per batch, an inclusion proof consists of approximately 20 SHA-256 hashes (640 bytes). The storage requirement for compliance with 8-year NHS retention, at the scale of an NHS trust processing 100,000 AI interactions per day, is approximately 18 MB per year in Merkle roots and proof anchors, manageable on-premises storage at any classification level.

Classification-Aware Deployment

The deployment configuration varies by classification level. For OFFICIAL and OFFICIAL-SENSITIVE systems, the anchoring service can be a public blockchain or a government cloud-hosted RFC 3161 service. For SECRET and TOP SECRET systems, the anchoring must be within the government's own infrastructure (PSNA or equivalent), and the ML-DSA-65 signing key must be held in a Hardware Security Module (HSM) that meets the relevant NCSC key management standards. The proof verification process is identical across classification levels; only the anchoring destination and key custody requirements differ.

SECTION 10

Data Residency and the Government Security Classification

UK government AI audit infrastructure must satisfy data residency requirements that do not apply to commercial deployments. The Government Security Classification (GSC) framework applies to all data handled by or on behalf of HM Government. AI audit records (even though they do not contain the content of AI interactions) may themselves be classified if the fact of an AI interaction at a specific time with a specific government system is itself sensitive information.

The residency requirements by classification level are:

CLASSIFICATION	RESIDENCY REQUIREMENT	CLOUD OPTIONS	ANCHORING REQUIREMENT
OFFICIAL	UK jurisdiction preferred; public cloud with Cyber Essentials Plus and ISO 27001 acceptable	Hyperscaler UK regions (AWS London, Azure UK South, GCP London); GovCloud	Public blockchain or commercial RFC 3161 acceptable
OFFICIAL-SENSITIVE	UK jurisdiction required; cloud must meet NCSC Cloud Security Principles (CSPs)	IL2-certified cloud (AWS GovCloud EU, Azure Government UK, G-Cloud framework)	Government-operated RFC 3161 or equivalent
SECRET	UK-controlled infrastructure; OFFICIAL cloud insufficient	IL3-accredited systems; MOD PACE or PSNA; commercial systems require specific accreditation	On-premises HSM anchoring; no public anchoring
TOP SECRET	Closed UK government network (STRAP-approved)	On-premises only; air-gapped where required	On-premises anchoring only

AffixIO's deployment model supports all four tiers. The cryptographic protocol is identical across tiers; the deployment environment, the network connectivity, and the key custody model adapt to classification requirements. This means a department deploying AI systems across multiple classification levels (OFFICIAL for public-facing services and SECRET for national security analysis) can use a single audit framework with consistent evidential properties across the full classification spectrum.

One specific requirement that distinguishes government AI audit from commercial deployments is support for the UK GDPR right of access (Article 15). An individual whose affairs were affected by an AI-assisted government decision has the right to request information about the automated processing that was applied. The audit trail architecture must be able to respond to Subject Access Requests without disclosing other individuals' interaction records. Merkle inclusion proofs support selective disclosure: the record for a specific interaction can be provided to the relevant individual without exposing the full batch of records from the same time period.

SECTION 11

From Pilot to Sovereign Standard

The UK's AI governance challenge in 2026 is not a shortage of pilots. The Incubator for AI (i.AI) has supported dozens of AI pilots across central government, from HMRC to the Home Office to the NHS. The challenge is scaling from pilot to sovereign standard: deploying AI systems at production scale, across sensitive data, with the accountability and auditability that parliamentary scrutiny, judicial review risk, and public trust require.

Three transitions are necessary to move from pilot to sovereign standard.

From Departmental to Cross-Government Accountability

Individual departments are currently developing their own AI audit approaches independently. HMRC's Quantexa deployment, the NHS AI Lab's clinical AI programmes, and i.AI's pilots each have different audit mechanisms, different log formats, and different accountability chains. A sovereign

standard requires interoperability: the ability for a cross-departmental audit (or a joint Parliamentary committee investigation) to retrieve consistent, comparable AI accountability evidence from multiple departments. This requires a common record format and a common anchoring protocol, not a common platform. The proof format and the Merkle anchoring protocol can be standardised without requiring departments to use the same AI systems or the same infrastructure.

From Pre-Deployment to Continuous Assurance

The current assurance model (CDDO framework compliance, AISI evaluation, DTAC review) is front-loaded. Assurance happens before deployment; the assumption is that a system that passed assurance will continue to behave within its assessed parameters. This assumption fails as AI systems are updated, fine-tuned, integrated with new data sources, and deployed in contexts that differ from those assessed. Continuous assurance requires runtime evidence, not just pre-deployment evidence. The cryptographic audit trail is the mechanism for continuous assurance; the AISI evaluation establishes the baseline against which runtime evidence is compared.

From Domestic Compliance to International Standard-Setting

The UK has an opportunity, not yet realised, to set the international standard for cryptographic AI accountability in government. The EU AI Act's Article 12 logging requirements are process-oriented rather than cryptographically specified; they require "logging" but do not specify what makes a log valid evidence. The UK's pro-innovation regulatory approach has created space to define a higher, more technically rigorous standard than the EU's process-based requirements. A UK standard that specifies Merkle-anchored, post-quantum-signed AI interaction records (with a clear protocol for cross-border verifiability) would give UK AI exports a demonstrable governance advantage in markets where AI accountability is increasingly mandated. AISI's existing international evaluation partnerships with the US, Japan, Singapore, and South Korea provide the diplomatic infrastructure through which such a standard could be promoted.

SECTION 12

Conclusion: The Audit Trail as National Infrastructure

The UK's sovereign AI programme is, in the main, well designed and well-funded. The investment in compute, the Sovereign AI Fund, the National Data Library, and the delivery progress on the AI Action Plan represent a serious and credible commitment to domestic AI capability. The CDDO framework, the AISI evaluation programme, the algorithmic transparency standard, and the NCSC's zero trust guidance provide a governance architecture that is more comprehensive than most comparable national frameworks.

The gap is accountability at runtime. The current governance architecture is excellent at establishing what AI systems should do, evaluating whether they meet safety thresholds before deployment, and documenting their intended purpose and data use. It is not yet able to answer, from within UK sovereign infrastructure, what a specific AI system actually did at a specific moment, whether it was operating within its authorised parameters at that moment, and whether the evidence of that behaviour is tamper-evident and judicially durable.

This is not a failure of ambition; it is the natural order of infrastructure development. Compute precedes accountability, because you cannot account for AI decisions until you have AI systems making them. The UK is now at the inflection point where AI systems are moving from pilot to production at scale (HMRC at tax collection scale, NHS at clinical-decision scale, DWP at benefits-determination scale) and the accountability infrastructure must scale with them.

The case for treating the AI audit trail as national infrastructure rests on four arguments:

- **Democratic legitimacy:** Parliament cannot scrutinise AI-assisted government decisions without verifiable records of those decisions. Mutable logs held by foreign technology vendors do not provide Parliament with that scrutiny capability.
- **Judicial protection:** Citizens subject to AI-assisted government decisions have rights under the Human Rights Act 1998 and judicial review that

require those decisions to be documentable and challengeable. The audit trail is the mechanism for that documentation.

- **National security:** Defence and national security AI must generate records that remain tamper-evident over decades and against adversaries with quantum computing capability. Post-quantum audit infrastructure is as much a national security requirement as post-quantum communications.
- **Economic advantage:** UK AI systems that can demonstrate cryptographic accountability will have a genuine competitive advantage in procurement by governments, regulated industries, and international organisations that require AI accountability as a condition of deployment.

AffixIO's position within this architecture is as a domestic UK provider of the proof layer: the cryptographic attestation infrastructure that makes AI interactions tamper-evidently attributable and verifiable, compatible with UK data residency requirements, the GSC framework, UK GDPR, and NCSC security standards. We are one component in a broader ecosystem that includes AI system providers, cloud infrastructure operators, independent auditors, and the government's own assurance functions. The audit trail is not the whole of sovereign AI governance. It is the part that makes everything else verifiable.

> Related AffixIO whitepapers: WP-034: Shadow AI Governance covers the enterprise shadow AI context that intersects with public sector deployment. WP-032: Sublinear Post-Quantum Attestation provides the detailed Merkle+ML-DSA-65 storage architecture. WP-002: Post-Quantum Attestation in Production with ML-DSA-65 covers the cryptographic foundations. WP-003: The Proof-Not-Log Paradigm addresses the evidentiary argument for cryptographic records over mutable logs.

FREQUENTLY ASKED

UK Sovereign AI Governance: Common Questions

What is the UK sovereign AI programme and what has it committed to?

The UK Sovereign AI programme is coordinated by DSIT and includes a £500 million Sovereign AI Fund (April 2026), a £750 million national AI supercomputer, a £600 million Health Data Research Service for NHS data access, and 38 of 50 AI Opportunities Action Plan commitments delivered by January 2026, tracked at delivery.ai.gov.uk.

What does the CDDO Generative AI Framework require for UK government AI systems?

The CDDO framework's accountability and transparency principles require that AI-assisted decisions are auditable, that the AI system version used in any decision is attributable, and that active safeguards were demonstrably in place at the moment of each decision. A mutable SIEM log does not satisfy these requirements. Tamper-evident, cryptographically signed interaction records do.

How does the AI Security Institute relate to ongoing AI governance?

AISI conducts pre-deployment evaluations of frontier AI models. Pre-deployment testing is a point-in-time assessment; it does not prove continuous compliance post-deployment. The runtime audit trail bridges this gap: it provides continuous, tamper-evident evidence that the live deployed system is behaving consistently with its evaluated baseline, enabling AISI evaluation to serve as the accountability anchor it is intended to be.

What does zero trust mean for AI governance in UK government?

Applied to AI, the NCSC's eight zero trust principles require explicit authorisation at the point of each AI interaction (not just at login), continuous monitoring against authorised scope, cryptographic evidence that each interaction was authorised and attributable, and records that are verifiable independently of the custodian holding them.

WP-034: Shadow AI Governance | WP-032: Sublinear Attestation | WP-003: Proof Not Log

▶ About
▶ Solutions
▶ Legal
▶ Trust & Security

Contact

truth layer | yes | no | proof