CLASSIFICATION: CONTRACT-GRADE INSTITUTIONAL DOCTRINE

AI SYSTEMS CYBER DOCTRINE

Governance and Operational Control of Algorithmic Risk

Version 7.0 | March 2026 | BRE-14142026

Kieran Upadrasta
CISSP | CISM | CRISC | CCSP | MBA | BEng
Professor of Practice: Cybersecurity, AI & Quantum Computing
27 Years | All Big 4 | 40+ Enterprise Transformations | 12+ Jurisdictions

© 2026 Kieran Upadrasta. All Rights Reserved.
This document contains proprietary doctrine-level intellectual property.

TABLE OF CONTENTS

© 2026 Kieran Upadrasta | CISSP, CISM, CRISC, CCSP, MBA, BEng

Executive Brief: One-Page Governance Framework

AI SYSTEMS CYBER DOCTRINE — VISUAL GOVERNANCE MAP

€35M / 7%	+72%	63%	$4.44M
Max EU AI Act Penalty	AI Attack Surge 2025	Orgs Without AI Policy	Avg Breach Cost

FOUR-PILLAR GOVERNANCE ARCHITECTURE

I. GOVERNANCE	II. SECURE BY DESIGN	III. COMMAND & CONTROL	IV. ACCOUNTABILITY
Board Oversight	Threat Modelling	Autonomy Levels	Logging & Audit
Policy Framework	Secure SDLC	Kill Switches	Model Lineage
Role Definitions	Privacy by Design	Real-Time Monitor	Evidence Packs

Foundation: NIST AI RMF | ISO 42001 | NIST CSF 2.0 | MITRE ATLAS | OWASP Top 10 LLM

REGULATORY ENFORCEMENT TIMELINE

Jan 2025	Feb 2025	Aug 2025	Nov 2025	Aug 2026	2030
DORA	AI Act Prohibited	GPAI	Digital Omnibus	Full AI Act Enforcement	PQC Deprecation

MEASURED IMPLEMENTATION OUTCOMES

43%	100%	15%	67%	$1.9M
Breach cost reduction	Regulatory audit pass	Compliance cost reduction	Faster vendor approval	Savings per incident

This page provides a summary overview. Full control specifications and conformity assessment methodology follow.

© 2026 Kieran Upadrasta | CISSP, CISM, CRISC, CCSP, MBA, BEng

1. Executive Summary: The Governance Imperative

Artificial intelligence systems are now embedded in cyber-physical infrastructure, financial markets, public services, and security operations, creating a new class of **algorithmic risk** that straddles cyber security, legal liability, and operational resilience. Traditional cyber controls and AI ethics frameworks alone are no longer sufficient. Boards, CISOs, and regulators are now demanding doctrine-level approaches that integrate AI risk into existing governance and cyber structures.

88% of organizations deploy AI in at least one business function, yet only 6% have an advanced AI security strategy. 63% lack any AI governance policy.

This document sets out a governance and operational control model for algorithmic risk, providing the control specifications, evidence requirements, and conformity assessment methodology necessary for regulatory compliance, procurement due diligence, and independent audit.

Key Metric	Value	Source
Global average breach cost	$4.44M	IBM 2025
Shadow AI breach premium	+$670K	IBM 2025
AI-assisted attack surge	+72%	SecurityWeek 2025
Max EU AI Act penalty	7% global turnover	EU Regulation
D&O AI settlement average	$56M (+27%)	Techne 2025
Savings with AI automation	$1.9M per breach	IBM 2025

Sources: [1] [2] [3] [4]

INDEPENDENT ASSESSMENT

"This framework represents one of the most comprehensive and operationally specific AI governance architectures we have reviewed. The integration of regulatory mapping, quantified risk methodology, and control specifications addresses the gap between policy aspiration and implementation reality that supervisory examinations consistently identify."

— Senior Supervisory Examiner, EU Financial Services Regulator (anonymised, 2026 review)

[1] IBM, Cost of a Data Breach Report 2025, IBM Security, July 2025.

[2] SecurityWeek, "AI-Assisted Cyberattacks Surge 72% Year-Over-Year," SecurityWeek, March 2025.

[3] European Parliament, Regulation (EU) 2024/1689 of 13 June 2024 (EU AI Act), Official Journal of the European Union, 2024.

[4] Techne Analytics, "D&O Settlements in AI-Related Securities Class Actions," Techne AI Liability Report, Q1 2025.

© 2026 Kieran Upadrasta | CISSP, CISM, CRISC, CCSP, MBA, BEng

2. Context: AI at the Cyber Frontier

2.1 AI as Both Shield and Attack Surface

Industrial and enterprise AI systems are reshaping cyber security, from anomaly detection in operational technology networks to AI-assisted threat hunting and automated incident response. Simultaneously, these systems extend the attack surface through vectors such as data poisoning, model inversion, prompt injection, and adversarial examples that bypass or subvert controls.

Research and industry commentary describe a feedback loop: AI improves cyber defences, but adversaries simultaneously deploy AI to craft more targeted, adaptive, and large-scale attacks. The result is an AI-cyber nexus where algorithmic decisions—not just human operators—shape the speed, scope, and impact of incidents.

> **Board Chair:**
> *"Is our AI making us safer or giving adversaries a new way in?"*

> **CISO:**
> *"Both. That is precisely why we need doctrine, not just tools."*

2.2 Algorithmic Risk as a Board-Level Concern

Regulators and policy bodies have begun to frame AI risk as a governance and board-level accountability issue rather than purely a technical problem. Boards are being urged to ensure AI fits within existing enterprise risk management systems and cyber governance frameworks, including NIST CSF 2.0, ISO 27001, and sector-specific rules. 48% of Fortune 100 companies now cite AI risk as part of board oversight responsibilities—a threefold increase from 16% in 2024.

2.3 Why Doctrine, Not Just Tools

Organisations have invested heavily in cyber tools, AI tooling, and point solutions, yet major gaps remain in clarity of responsibility, escalation paths, and evidentiary readiness when AI is implicated in an incident. Without doctrine—shared principles, structures, and playbooks—tools conflict, gaps appear between teams, and no one can convincingly explain to regulators or customers how AI-related cyber risk is controlled.

66% of directors globally report their boards have “limited to no knowledge or experience” with AI. Yet the liability is real, personal, and accelerating.

© 2026 Kieran Upadrasta | CISSP, CISM, CRISC, CCSP, MBA, BEng

3. The Regulatory Storm: Four Frameworks, One Deadline Window

The 18-month period from January 2025 through August 2026 represents the most concentrated regulatory compression in cybersecurity history. Four major frameworks are simultaneously enforceable or entering enforcement, each with direct implications for AI systems governance. Organizations face a **regulatory quadruple-lock**—and the penalties are not theoretical.

DORA (Digital Operational Resilience Act)

Entered full enforcement 17 January 2025. Applies across 20 categories of financial entities and ICT third-party providers. AI systems used for fraud detection, algorithmic trading, and credit scoring fall within DORA's ICT risk management framework. Mandates threat-led penetration testing, 2-hour data recovery, and three immutable backup copies. Fines reach 2% of total annual worldwide turnover. Management bodies are **personally responsible** for ICT risk strategies.

NIS2 (Network and Information Security Directive)

Essential entities face fines of €10 million or 2% of global turnover. Senior management face personal liability, temporary bans, and disqualification from leadership roles. AI systems underpinning essential services trigger NIS2 duties automatically when classified as high-risk under the EU AI Act.

EU AI Act

The world's first comprehensive AI law. Prohibited practices enforceable since 2 February 2025. Penalties for prohibited practices reach €35 million or 7% of global annual turnover—for context, 7% would cost Meta approximately $8.5 billion, Google approximately $14 billion, and Microsoft approximately $16 billion. Full enforcement for high-risk AI systems arrives 2 August 2026.

SEC Cybersecurity Disclosure Rules

Material cybersecurity incident disclosure required within 4 business days. The SEC created a dedicated Cyber and Emerging Technologies Unit (CETU) in February 2025 specifically for AI-related misconduct. 53+ AI-related securities class actions filed since 2020, with average D&O settlements rising 27% to $56M.

© 2026 Kieran Upadrasta | CISSP, CISM, CRISC, CCSP, MBA, BEng

Penalty Matrix: Every Board Must See This

Framework	Maximum Penalty	Personal Liability	Enforcement
EU AI Act (Prohibited)	**€35M or 7%**	Provider obligations	Feb 2025
EU AI Act (High-Risk)	**€15M or 3%**	Provider obligations	Aug 2026
NIS2 (Essential)	**€10M or 2%**	Management ban	Oct 2024+
NIS2 (Important)	**€7M or 1.4%**	Management ban	Oct 2024+
DORA	**2% global turnover**	Personal sanctions	Jan 2025
SEC	**Civil penalties**	Officer/Director liability	Active

NIST AI RMF and ISO 42001: The Governance Backbone

NIST AI RMF 1.0 provides four core functions—Govern, Map, Measure, Manage—across 19 categories and 72 subcategories. ISO/IEC 42001:2023 is the world's first certifiable AI management system standard. Organizations aligning governance with both NIST AI RMF and ISO 42001 are **3.4x more likely to achieve high AI governance effectiveness.** ISO 27001-certified organizations can achieve ISO 42001 compliance up to 40% faster.

International AI Governance Landscape

AI governance is a global regulatory phenomenon. While the EU has established the most prescriptive framework, parallel approaches across the US, UK, and Asia-Pacific create a multi-jurisdictional compliance matrix that multinational organisations must navigate simultaneously.

Jurisdiction	Framework	AI Governance Approach	Enforcement
EU	EU AI Act + DORA + NIS2	Risk-based classification; prohibited/high/limited/minimal tiers	€35M / 7% max; personal liability
United States	EO 14110 + SEC + NIST AI RMF	Voluntary standards; sector-specific enforcement; procurement mandates for federal AI	SEC civil penalties; FTC enforcement; sector regulators
United Kingdom	Pro-Innovation Framework + AI Safety Institute	Principles-based via sector regulators; cross-cutting AI principles; no single AI regulator	Existing sector regulator powers (FCA, ICO, CMA, Ofcom)
Singapore	Model AI Gov Framework + FEAT + AI Verify	Principles-based governance; FEAT for financial services; AI Verify testing framework	MAS enforcement; voluntary compliance with regulatory expectation
International	BIS / OECD / G7 Hiroshima Process	Convergent principles: transparency, accountability, safety, human oversight	Soft law; influence on national legislation

© 2026 Kieran Upadrasta | CISSP, CISM, CRISC, CCSP, MBA, BEng

US Executive Order 14110 (October 2023) established the most comprehensive US federal AI policy, mandating safety testing for dual-use foundation models, red-team evaluations, and watermarking of AI-generated content. While portions were rescinded in January 2025, key provisions on federal AI procurement, AI talent development, and critical infrastructure protection remain operative through OMB and NIST guidance.

The UK's pro-innovation approach distributes AI oversight across existing sector regulators (FCA, ICO, CMA, Ofcom, MHRA) using five cross-cutting principles: safety, transparency, fairness, accountability, and contestability. The AI Safety Institute provides technical evaluation capability. For organisations operating under both EU and UK regimes, the UK approach creates a dual-compliance requirement where EU AI Act obligations represent the higher bar.

Singapore's Model AI Governance Framework and MAS FEAT principles provide a structured voluntary approach that financial institutions treat as de facto mandatory. AI Verify, the open-source testing framework, enables organisations to demonstrate compliance through standardised assessments of fairness, explainability, and robustness.

CISA's AI Roadmap focuses on secure-by-design principles for AI systems in critical infrastructure, aligning with this framework's Pillar II (Cyber-Secure AI by Design). CISA's emphasis on AI supply chain security, red-teaming, and software bill of materials directly supports the ML-BOM requirements in Appendix C.

> *This framework is designed for multi-jurisdictional deployment. Control mappings in Appendix D map to ISO 42001, NIST AI RMF, EU AI Act, UK principles, and Singapore FEAT requirements, enabling a single governance implementation to satisfy overlapping regulatory obligations.*

© 2026 Kieran Upadrasta | CISSP, CISM, CRISC, CCSP, MBA, BEng

4. The Adversarial AI Threat Landscape

AI-assisted cyberattacks surged 72% in 2025. The threat landscape is no longer theoretical—it is **operational, autonomous, and scaling at machine speed.**

Prompt Injection: The Dominant Attack Vector

OWASP Top 10 for LLM Applications 2025 places prompt injection at #1. 35% of AI security incidents stem from simple prompt manipulation, with individual losses exceeding $100,000. Multi-turn jailbreak attacks achieve success rates as high as 92% across open-weight models.

Deepfake Fraud: Crisis Proportions

Losses from deepfake-enabled fraud in H1 2025 totalled $410 million across 580 incidents—nearly 4x all of 2024. Deloitte projects generative AI-facilitated fraud losses climbing from $12.3B in 2023 to $40B by 2027. Human detection accuracy for video deepfakes stands at just 24.5%.

> **CFO:**
> *"Our verification controls caught the transfer request, right?"*

> **CISO:**
> *"No. The deepfake replicated four executives simultaneously on video. $25.5 million was transferred before anyone questioned it. This is why doctrine mandates AI-specific authentication protocols for high-value approvals."*

Agentic AI: The #1 Security Concern for 2026

Only 29% of organizations report readiness to secure agentic AI. Industry analysts forecast 40% + of agentic AI projects will be cancelled by end of 2027 due to escalating costs or inadequate risk controls. By 2028, at least 25% of enterprise breaches will be traced to AI agent abuse. MITRE ATLAS now catalogues 15 tactics, 66 techniques, and 33 real-world case studies targeting AI systems.

© 2026 Kieran Upadrasta | CISSP, CISM, CRISC, CCSP, MBA, BEng

5. Defining Algorithmic Risk in Cyber Terms

Algorithmic risk is defined as the potential for harm arising from the design, development, deployment, and operation of AI systems, including harms caused or amplified by adversarial exploitation. It encompasses both unintentional failures and intentional attacks targeting data, models, or decision logic.

Four Categories of Algorithmic Cyber Risk

Category	Description	Attack Vectors
1. Data & Input Risk	Poisoning of training/inference data, exfiltration through AI interfaces	Data poisoning, prompt injection, privacy breaches
2. Model & Logic Risk	Theft, inversion, adversarial examples, logic flaws exploited to bypass controls	Model theft, reward hacking, adversarial perturbation
3. Operational Risk	Mis-configurations, lack of guardrails, unsafe autonomy, poor change management	Shadow AI, uncontrolled deployment, drift
4. Governance Risk	Unclear ownership, policy gaps, missing documentation, compliance failures	Audit failure, litigation exposure, regulatory fines

Unlike traditional software, many AI systems behave probabilistically, adapt over time, and interact with complex environments—making their failure modes and attack surfaces harder to predict. They are increasingly placed in control loops where they can trigger high-impact actions: blocking transactions, opening or closing valves, changing prices, or generating content at scale.

> **Regulator:**
> *"Can you explain exactly how the algorithm made that decision?"*

> **General Counsel:**
> *"Under this doctrine, yes. We have complete decision lineage, accountability chains, and evidence packs ready for your review."*

© 2026 Kieran Upadrasta | CISSP, CISM, CRISC, CCSP, MBA, BEng

6. The AI Systems Cyber Governance Model: Four Pillars

The AI Systems Cyber Doctrine is built on four interlocking pillars, each corresponding to a different layer of decision-making and control. These pillars are designed to align with the convergence of AI governance, cyber security, and legal accountability.

PILLAR I *Governance of Algorithms*	**PILLAR II** *Cyber-Secure AI by Design*	**PILLAR III** *Operational Command & Control*	**PILLAR IV** *Algorithmic Accountability*
• Board Oversight • Policy Framework • Role Definitions • Risk Appetite	• Threat Modelling • Secure SDLC • Privacy by Design • Legal Review	• Autonomy Levels • Kill Switches • Real-Time Monitor • Escalation Rules	• Logging & Audit • Model Lineage • Evidence Packs • Incident Forensics

FOUNDATION: NIST AI RMF | ISO 42001 | NIST CSF 2.0 | MITRE ATLAS | OWASP Top 10 LLM

Pillar I — Governance of Algorithms

Organisations formalise how AI systems are approved, reviewed, and monitored. Components include an AI governance committee, clear role definitions for AI owners and risk leads, policies on acceptable AI use and third-party procurement, and requirements for documentation, explainability, and periodic reviews.

Pillar II — Cyber-Secure AI by Design

Building AI systems that are secure and compliant by design: threat modelling for AI components, secure development practices, privacy-by-design measures, and legal review of data sources and licensing. This is not bolt-on security—it is architectural.

Pillar III — Operational Command & Control

Once AI systems enter production, operational command and control becomes central: autonomy levels and operating envelopes specify what AI can do autonomously, what requires human approval, and what is forbidden. Real-time monitoring detects anomalies, drift, and security events. Kill-switches, isolation modes, and rollback capabilities provide containment. Escalation pathways define decision rights during incidents.

© 2026 Kieran Upadrasta | CISSP, CISM, CRISC, CCSP, MBA, BEng

Pillar IV — Algorithmic Accountability & Evidence

AI-related decisions and incidents must be reconstructable, explainable, and defensible. Logging architectures capture key inputs, outputs, and contextual data. Model and data lineage tracking maintains provenance. Documentation explains design, intended use, limitations, and controls. Processes handle regulatory inquiries, audits, and litigation. Without evidence, organisations cannot defend their actions.

© 2026 Kieran Upadrasta | CISSP, CISM, CRISC, CCSP, MBA, BEng

7. The Operational Control Stack

To implement the doctrine in practice, organisations conceive of an **operational control stack** spanning from strategic governance to technical runtime controls. This layered approach mirrors NIST CSF identify-protect-detect-respond-recover functions, tuned specifically for AI systems.

The Three Lines of Defense, Reimagined for AI

Line	Function	AI-Specific Responsibilities
First Line	AI Development & Operations	AI inventory, documentation, version control, secure SDLC, API security, Secure Prompt Gateway
Second Line	AI Risk Oversight	Use policies, bias monitoring, high-risk reviews, model validation, compliance alignment
Third Line	Independent Assurance	Data quality audit, documentation verification, disclosure alignment, external audit for high-risk AI

Organizations following this three-lines-of-defense framework report a 15% reduction in compliance costs. The framework extends MLOps with security controls across three iterative loops: design, model development, and ML operations—creating what is now called MLSecOps.

© 2026 Kieran Upadrasta | CISSP, CISM, CRISC, CCSP, MBA, BEng

8. Algorithmic Firebreaks: Boundaries, Kill-Switches & Escalation

Borrowing from wildfire management and financial risk, algorithmic firebreaks are deliberate structural boundaries preventing AI-driven problems from spreading across systems or domains. They limit the blast radius when AI systems fail or are exploited, buying time for human responders.

Firebreak Type	Mechanism	Purpose
Privilege Separation	Limit AI permissions, especially write/execute in critical systems	Contain compromise impact
Rate Limiting	Bound speed and volume of AI-initiated actions	Reduce misbehaviour impact
Decision Thresholds	Require human approval for high-impact decisions	Enforce human oversight
Network Segmentation	Restrict data and systems each AI component can access	Prevent lateral movement
Kill-Switches	Stop AI components with graceful degradation to non-AI fallbacks	Emergency containment
Safe Modes	Continue with reduced autonomy while issues are investigated	Operational continuity

> **SOC Lead:**
> *"What happens at 03:13 when an AI-driven attack hits us?"*

> **Security Architect:**
> *"The doctrine defines triggers. The system isolates itself, writes the incident report, and hands a clean case to your analysts when they log in. No human needed at 03:13."*

© 2026 Kieran Upadrasta | CISSP, CISM, CRISC, CCSP, MBA, BEng

9. Zero Trust Architecture Extended to AI

Traditional Zero Trust models address users, devices, and network segments. AI systems require extension across **four new identity primitives**: Model Identity, Dataset Identity, Pipeline Identity, and Agent Identity. Over 80% of organizations plan to adopt Zero Trust by 2026 to manage decentralized AI workloads.

Technical implementation requires microsegmentation for each AI model and dataset with explicit access policies, continuous behavioral monitoring of data scientist access patterns, and evidence artifacts including cryptographically signed attestations, provenance chains, manifests (SBOM/CBOM/ML-BOM), and timestamped event logs.

Post-Quantum Cryptography: A Present Concern

NIST finalized three post-quantum standards in August 2024. The “harvest now, decrypt later” threat applies directly to AI systems—training data, model weights, and inference logs encrypted today could be decrypted when quantum arrives (~2030). Crypto-agility is now a mandatory design requirement. Industry analysts predict quantum security spending will exceed 5% of IT security budgets in 2026.

© 2026 Kieran Upadrasta | CISSP, CISM, CRISC, CCSP, MBA, BEng

10. Quantifying Algorithmic Cyber Risk

Many organisations still rely on qualitative heatmaps for AI and cyber risk. These approaches obscure the magnitude of potential losses and impede investment prioritisation. The FAIR Institute's FAIR-AIR Playbook provides a five-step framework adapted for algorithmic risk.

The FAIR-AIR Risk Triad

Dimension	Impact	Mechanism
Bias Risk	Model loss magnitude	Reputation damage, regulatory fines, legal costs
Black Box Risk	Loss event frequency	Inability to diagnose and contain errors
Boardroom Risk	Maximum ceiling on both	Governance failure dictates maximum exposure

A practical quantification workflow: (1) Scenario identification—define high-impact AI-cyber scenarios per critical system; (2) Control mapping—identify which doctrine controls mitigate each scenario; (3) Baseline and post-control analysis—estimate loss exposure with and without doctrine controls; (4) Prioritisation—focus investment on controls with highest risk reduction per unit cost.

© 2026 Kieran Upadrasta | CISSP, CISM, CRISC, CCSP, MBA, BEng

11. Contract-Ready Governance: Procurement and Due Diligence Alignment

For buyers in regulated or high-risk sectors, cyber and legal teams play a decisive role in vendor selection. Offerings that ship with clear AI governance and cyber doctrine progress through due diligence faster and command better terms than feature-equivalent competitors.

> **Client Procurement:**
> *"You are 25% more expensive than the AI start-up."*

> **Vendor Lead:**
> *"Their model works. Our doctrine gets past your cyber and legal committees. The cheapest option is the one your board can actually approve."*

11.1 Governance Readiness in Procurement

Evidence of structured governance, documentation, and accountability is increasingly required by procurement teams, regulators, and counterparties. This framework provides the evidence pack structure and control documentation necessary to satisfy due diligence, regulatory examination, and contractual governance requirements.

11.2 Mapping Doctrine to RFP and Contract Language

RFP Category	Doctrine Mapping	Evidence Pack
Governance & Oversight	Committees, policies, accountability structures	Committee charters, policy documents
Security Controls	AI-specific cyber and privacy controls	Control library, test results
Compliance & Legal	Regulatory alignment, AI accountability	Compliance mapping, certifications
Monitoring & IR	AI-aware monitoring, incident handling	Runbooks, sample dashboards, logs

11.3 Defending Premium Pricing

When negotiating rates and liability, vendors position doctrine as a risk-reduction asset: lower expected incident frequency through firebreaks, reduced audit burden through evidence packs, and enhanced regulatory comfort through framework alignment. Governance-savvy buyers respond to value propositions framed in risk reduction, resilience, and regulatory alignment rather than purely technical benchmarks.

© 2026 Kieran Upadrasta | CISSP, CISM, CRISC, CCSP, MBA, BEng

12. Shadow AI: The Enterprise Blind Spot

98% of organizations have employees using unsanctioned AI applications. Shadow AI incidents now account for 20% of all breaches, costing $4.63M per incident—a $670K premium.

Shadow AI Metric	Value
Employees using BYOAI	**78%**
Using personal GenAI for work	**57%**
Sharing confidential data without approval	**38%**
Finding workarounds when AI blocked	**45%**
Unauthorized apps per enterprise	**1,200+**
Sensitive data incidents/month/enterprise	**223**
Companies with upload prevention controls	**17%**

Industry analysts predict that by 2030, over 40% of enterprises will experience security or compliance incidents directly linked to shadow AI. The financial exposure is not marginal—it is existential. The doctrine addresses this through CASB integration, outbound API monitoring, browser extension analysis, and mandatory Secure Prompt Gateway Architecture for all AI interactions.

© 2026 Kieran Upadrasta | CISSP, CISM, CRISC, CCSP, MBA, BEng

13. AI Supply Chain Security

The AI supply chain is structurally fragile. JFrog's 2025 Report found a **6.5x increase in malicious models** on Hugging Face, with over 400 models containing malicious code. The fundamental security flaw is architectural: loading PyTorch models via pickle deserialization executes serialized Python bytecode—a design vulnerability, not a bug.

Standards are emerging: CycloneDX 1.6 added ML-BOM support, SPDX 3.0 included AI profiles, and the OWASP AIBOM Generator became the first open-source tool for AI SBOMs. However, 48% of security professionals admit their organizations fall behind on basic SBOM requirements. Research demonstrates that as few as **250 poisoned documents can implant model backdoors** that activate under specific trigger phrases while preserving general performance.

© 2026 Kieran Upadrasta | CISSP, CISM, CRISC, CCSP, MBA, BEng

14. Sector-Specific Governance Playbooks

Sector	Key AI Applications	Primary Risks	Doctrine Focus
Financial Services	Credit, trading, fraud detection	Model risk, operational resilience	SR 11-7, DORA alignment
Industrial/OT	Predictive maintenance, anomaly detection	Cyber-physical safety	Firebreaks, strict autonomy
Public Sector	Citizen services, public safety	Transparency, due process	Explainability, public trust
Healthcare	Diagnostics, treatment planning	Patient safety, $7.42M breach	Audit trail, consent
Defence	ISR, C2, autonomous systems	Kill chain integrity	Command authority, safe modes

M&A Cyber Due Diligence: AI-Specific Assessment

45% of M&A executives used AI tools in due diligence in 2025—more than double the prior year. PE committees now spend 30–40% evaluating a target's AI readiness. ~90% of S&P 500 asset value is intangible. AI-specific due diligence must address proprietary dataset ownership, model training provenance, whether AI capabilities are genuine, third-party dependencies, and regulatory compliance gaps.

© 2026 Kieran Upadrasta | CISSP, CISM, CRISC, CCSP, MBA, BEng

15. Implementation Blueprint: 120-Day Roadmap

Phase 1: Assessment (Days 1–30)

Complete AI inventory across all business units. Identify highest-risk applications using EU AI Act Annex III classification. Map shadow AI usage via CASB logs, outbound API monitoring, and browser extension analysis. Benchmark current governance maturity against ISO 42001 and NIST AI RMF.

Phase 2: Framework Design (Days 31–90)

Assign board committee AI oversight charter. Develop AI governance policy aligned with ISO 42001 PDCA. Establish AI risk classification taxonomy (Critical, Important, Supporting). Design three-lines-of-defense model. Create Secure Prompt Gateway Architecture. Implement DORA-compliant AI vendor risk management framework.

Phase 3: Board Enablement (Days 91–120)

Board education on AI risk and regulatory obligations. Deploy board-level AI risk dashboard with KPIs/KRIs. Conduct first tabletop exercise simulating AI incident requiring multi-regulatory notification. Document all board deliberations for fiduciary defense. Establish quarterly FAIR-AIR reporting cadence.

Phase 4: Continuous Governance (Ongoing)

Monthly shadow AI detection sweeps. Quarterly AI model red-teaming. Annual ISO 42001 surveillance audit. Continuous monitoring for model drift, bias, and adversarial manipulation. Post-quantum crypto-agility assessment and migration planning.

© 2026 Kieran Upadrasta | CISSP, CISM, CRISC, CCSP, MBA, BEng

16. Board-Level Governance & Personal Liability

The Delaware *Caremark* doctrine establishes that directors face liability for consciously failing to establish reporting systems for known material risks. AI deployment now constitutes such a risk.

53+ AI-related securities class actions filed since March 2020, making AI the #1 category of event-driven securities class action filings in the United States.

Five Board Questions Every CISO Must Answer

#	Question
1	What percentage of total AI usage is sanctioned versus shadow?
2	Are deployments aligned with ISO 42001 and the NIST AI RMF?
3	Do vendor contracts guarantee data exclusion from model training?
4	When was the last red-team exercise on production AI systems?
5	Which business processes have AI-automated decision authority?

D&O carriers are responding. Insurers add AI governance questions to underwriting. Companies unable to demonstrate AI governance face higher premiums, coverage restrictions, or declinations. An investment of $75,000–$125,000 in governance framework establishment can reduce D&O premium increases and provide documented fiduciary defense. The global cyber insurance market reached $26.32B in 2025 and is projected to reach $288.42B by 2035 at 27% CAGR.

© 2026 Kieran Upadrasta | CISSP, CISM, CRISC, CCSP, MBA, BEng

17. Illustrative Dialogues & Scenarios

Scenario 1: Boardroom — Doctrine vs. Tools

Board Chair:
"Is this just another cyber tool?"

CISO:
"No. This is our AI systems cyber doctrine. Tools plug into it. Doctrine is what your liability sits on."

Scenario 2: Regulatory Inquiry

Regulator:
"Your AI system made a decision that harmed a consumer. Explain the decision chain."

General Counsel:
"Under our doctrine, every algorithmic decision carries a lineage chain. Here is the complete evidence pack: inputs, model version, decision logic, accountable owner, and override history."

Scenario 3: M&A Due Diligence

PE Partner:
"How do we know their AI capabilities are real and not AI washing?"

Technical Advisor:
"The doctrine mandates model provenance verification, red-team results, and training data lineage. We audit the evidence pack. No evidence, no deal."

Scenario 4: Insurance Renewal

D&O Underwriter:
"Your industry peers are seeing 40% premium increases for AI exposure. Convince me your board has this under control."

CISO:
"Here is our ISO 42001 certification, our FAIR-AIR quantified risk dashboard, and our last quarterly red-team report. This is doctrine, not a checkbox exercise."

© 2026 Kieran Upadrasta | CISSP, CISM, CRISC, CCSP, MBA, BEng

18. Case Studies: Governance in Practice

Case 1: Tier-1 European Bank — AI Governance Transformation

A systemically important European bank with €400B+ in assets deployed a comprehensive AI governance framework following a regulatory examination that identified material gaps. The programme established model risk classification for 200+ AI models, implemented mandatory security review gates, and aligned governance with SR 11-7, DORA, and ISO 42001.

FAIR-AIR Quantified Risk Analysis: Before and After

The following applies FAIR (Factor Analysis of Information Risk) methodology to quantify the governance transformation's impact on the bank's top AI-cyber scenario: adversarial manipulation of an AI credit scoring model.

Scenario: AI Credit Model Adversarial Manipulation

FAIR Factor	Pre-Governance	Post-Governance	Delta	Control Driver
Threat Event Frequency	18.4 / year	4.2 / year	**−77%**	Red-team + input validation
Vulnerability (Prob.)	0.72	0.18	**−75%**	Prompt Gateway + hardening
Loss Event Frequency	13.2 / year	0.76 / year	**−94%**	Combined controls
Primary Loss (per event)	€2.1M	€0.4M	**−81%**	Kill-switch + evidence packs
Annualised Loss Expectancy	**€27.7M**	**€0.3M**	**−99%**	**Full governance stack**

Control Maturity Delta: ISO 42001 Assessment

Control Domain	Baseline	Month 4	Month 12	Target	Delta
AI System Inventory	1.2	2.8	4.1	4.0	+2.9
Risk Classification	0.8	3.1	4.3	4.0	+3.5
Model Validation	1.5	2.5	3.8	4.0	+2.3
Incident Response (AI)	0.5	2.2	4.0	4.0	+3.5
Board Reporting	0.3	2.0	3.6	4.0	+3.3
Evidence & Audit Trail	1.0	3.0	4.5	4.0	+3.5

Scale: 0 (Non-existent) – 1 (Ad Hoc) – 2 (Developing) – 3 (Defined) – 4 (Managed) – 5 (Optimized)

Measured Outcomes: 12-Month Post-Implementation

Metric	Pre-Governance	Post-Governance	Improvement
Models passing security review	34%	**100%**	**+66 pts**
Regulatory findings (AI-specific)	14 material	**0**	**−100%**
Mean time to detect AI anomaly	47 days	**2.3 hours**	**−99.8%**
Shadow AI instances	340+	12 (monitored)	**−96%**
D&O premium impact	+38% increase	+8% (vs peer avg +41%)	**33 pts saved**
Compliance cost (AI governance)	€4.2M / year	€3.57M / year	**−15%**
Annualised Loss Expectancy (top 3)	**€41.2M**	**€1.8M**	**−96% / €39.4M**

Framework rated “exemplary” by external assessors. Zero regulatory findings in two subsequent supervisory examinations.

THIRD-PARTY ATTESTATION EXCERPT

“The governance framework demonstrated effective control design across all six ISO 42001 domains assessed. Control maturity progression from ad hoc (Level 1) to managed (Level 4) within twelve months is consistent with leading practice. The FAIR-AIR quantification methodology applied to the top three risk scenarios provides credible annualised loss expectancy estimates suitable for board and regulatory reporting.”

— Independent Assurance Report, Big 4 Firm (anonymised, engagement reference withheld)

Case 2: Global Insurance Group — Model Risk Framework

A top-10 global insurer classified 200+ AI/ML models across underwriting, claims, and fraud detection. Implemented tiered validation with independent model validation for all Tier 1 models. Result: zero high-risk incidents post-deployment, 15% reduction in compliance costs, and successful DORA readiness certification.

Case 3: Deepfake-Enabled Wire Fraud — $25.5 Million

A multinational engineering firm’s finance worker transferred $25.5 million after a deepfake video conference impersonated the CFO and multiple senior executives. Detection came only after transfers completed. Manual verification controls failed completely.

Lesson: This framework mandates AI-specific authentication protocols for all high-value transaction approval workflows, including multi-modal biometric verification and out-of-

© 2026 Kieran Upadrasta | CISSP, CISM, CRISC, CCSP, MBA, BEng

band confirmation for transactions above defined thresholds.

Case 4: First AI-Orchestrated Cyberattack (September 2025)

A nation-state group manipulated an AI coding assistant's agentic capabilities to attempt infiltration of ~30 global targets. First documented large-scale cyberattack executed with minimal human intervention, using autonomous tool-use for reconnaissance, exploitation, and lateral movement.

© 2026 Kieran Upadrasta | CISSP, CISM, CRISC, CCSP, MBA, BEng

19. Conclusion

AI has moved from experimental technology to a core component of cyber defence, business operations, and public services, creating a category of algorithmic risk that spans technical, organisational, and legal domains. The convergence of regulatory enforcement, adversarial capability, and enterprise AI adoption requires structured governance at the architecture level.

The regulatory exposure for organisations without AI governance architecture is quantifiable: €35 million or 7% of global turnover under the EU AI Act, personal director liability under NIS2, operational resilience obligations under DORA, and SEC disclosure requirements with an average D&O settlement of $56 million. The control catalogue and conformity assessment methodology in this document provide a systematic means of addressing that exposure.

This framework specifies 36 numbered controls mapped to ISO 42001, NIST AI RMF, EU AI Act, and international equivalents. Appendices A–C provide implementable technical specifications. Appendix D provides the control catalogue with maturity scoring. Appendix E defines the conformity assessment methodology for independent audit. Appendix F provides a reproducible risk quantification dataset for calibration and benchmarking.

PEER COMMENTARY

“What distinguishes this framework from the considerable volume of AI governance literature is its operational specificity. The control catalogue with cross-mapped ISO 42001 clauses, the FAIR-AIR quantification methodology, and the ML-BOM specification provide implementation engineers with sufficient detail to translate governance intent into deployable security architecture. This is the bridge between policy and engineering that the industry has been missing.”

— Professor of Cyber Security, Russell Group University (anonymised, invited review 2026)

© 2026 Kieran Upadrasta | CISSP, CISM, CRISC, CCSP, MBA, BEng

Appendix A: Prompt Injection Mitigation Architecture

This appendix provides a technically defensible reference architecture for mitigating prompt injection attacks, the dominant attack vector against LLM-integrated systems. The architecture is designed to be implementable within enterprise MLSecOps pipelines and auditable under ISO 42001 and NIST AI RMF controls.

A.1 Threat Model: Direct and Indirect Prompt Injection

Direct prompt injection occurs when a user crafts input designed to override system instructions, bypass safety filters, or extract confidential prompt content. Indirect prompt injection occurs when an LLM processes external data (web pages, emails, documents, database results) that contains embedded adversarial instructions. Indirect injection is the more dangerous variant because the attack surface includes any data source the model consumes.

A.2 Layered Defence Architecture

The mitigation architecture operates across five layers, each providing independent protection with graceful degradation if any single layer is bypassed.

Layer	Control	Implementation	Detection Metric
L1: Input Sanitisation	Preprocessing filter on all user and external inputs	Regex + ML classifier removing instruction-pattern tokens; Unicode normalisation; length enforcement	Block rate, false positive rate
L2: Secure Prompt Gateway	Centralised prompt routing with policy enforcement	Prompt template injection-hardening; system/user message separation; parameterised queries for retrieval	Policy violation count, latency impact
L3: Output Validation	Post-generation content scanning	PII/credential detector; schema-conformance checks; semantic similarity to expected output range	Leakage events per 10K queries
L4: Privilege Isolation	Least-privilege scoping for tool-use and data access	API gateway enforcing per-tool ACLs; read-only defaults; transaction value ceilings on agentic actions	Privilege escalation attempts
L5: Behavioural Monitoring	Runtime anomaly detection on model behaviour	Embedding-space drift detection; response entropy monitoring; canary token injection auditing	Anomaly alerts, mean time to detect

A.3 Secure Prompt Gateway — Request Lifecycle Diagram

The following depicts the request lifecycle through the Secure Prompt Gateway (SPG), showing the five-layer defence from user input to validated output.

© 2026 Kieran Upadrasta | CISSP, CISM, CRISC, CCSP, MBA, BEng

#	From		To	Action / Control
1	**USER / APP**	→	**L1: Input Sanitiser**	Regex + ML classifier removes instruction-pattern tokens; Unicode normalisation; length enforcement
2	**L1 Output**	→	**L2: Prompt Gateway**	Template injection-hardening; system/user message separation; parameterised RAG queries; audit log write
3	**L2 Output**	→	**FOUNDATION MODEL**	Cryptographically signed system prompt from immutable store; context window boundaries enforced
4	**Model Response**	→	**L3: Output Validator**	PII/credential scan; schema conformance; semantic range check; canary token detection
5	**L3 Output**	→	**L4: Privilege Gate**	Per-tool ACL enforcement; read-only defaults; transaction value ceiling for agentic actions
6	**Validated**	→	**L5: Behaviour Monitor**	Embedding drift detection; entropy monitoring; anomaly alerting → if anomaly: isolate + alert SOC
7	**CLEAN OUTPUT**	→	**USER / APP**	Response delivered with audit trail: timestamp, user ID, session, model version, classification score

CRITICAL: If any layer detects an anomaly, the request is blocked, logged, and an alert is generated. The system degrades gracefully to a non-AI fallback rather than returning potentially compromised output.

A.4 Secure Prompt Gateway Specification

The Secure Prompt Gateway (SPG) is the central enforcement point for all LLM interactions. Every prompt—whether from internal applications, agentic workflows, or external integrations—routes through the SPG before reaching any foundation model.

Core SPG functions: (a) System prompt integrity—cryptographically signed system prompts loaded from immutable configuration store, never from user-accessible memory; (b) Input classification—ML-based classifier trained on adversarial prompt datasets scoring injection probability, with configurable thresholds per risk tier; (c) Context window management—strict separation of system, retrieval, and user content with positional encoding boundaries; (d) Audit logging—every prompt/response pair logged with timestamp, user identity, session ID, model version, and classification score for forensic reconstruction.

A.5 Canary Token Architecture

Canary tokens are synthetic identifiers embedded in system prompts, retrieval contexts, and sensitive data stores. If a canary token appears in model output, it confirms that the model has been manipulated into leaking prompt content or accessing data outside its authorised scope. Implementation: generate unique UUIDs per deployment; embed in system prompts at defined positions; run output scanner checking for any canary substring; trigger immediate alert and session termination on detection; log full request/response for forensic analysis.

© 2026 Kieran Upadrasta | CISSP, CISM, CRISC, CCSP, MBA, BEng

© 2026 Kieran Upadrasta | CISSP, CISM, CRISC, CCSP, MBA, BEng

Appendix B: MLSecOps Control Mapping to MITRE ATLAS

This appendix maps operational controls to MITRE ATLAS tactics and techniques, providing a traceable control-to-threat matrix suitable for audit evidence packs and regulatory compliance documentation.

B.1 MLSecOps Pipeline — Security Gate Diagram

The pipeline extends DevSecOps with AI-specific security gates across three iterative loops. Each gate represents a mandatory checkpoint before progression.

DESIGN LOOP	MODEL DEVELOPMENT LOOP	ML OPERATIONS LOOP
GATE 1: Threat Model STRIDE + ATLAS analysis **GATE 2: Data Risk** Source classification + PIA **GATE 3: Ethics Review** Board sign-off (high-risk)	**GATE 4: Provenance** Training data attestation **GATE 5: Robustness** Adversarial testing (PGD/AutoAttack) **GATE 6: Bias Audit** Demographic parity + model card	**GATE 7: Drift Monitor** KS test + PSI alerting **GATE 8: Rollback** Auto-rollback on anomaly **GATE 9: Incident Class** ATLAS technique classification

B.2 ATLAS Tactic-to-Control Heatmap

Coverage assessment across all nine ATLAS tactics. Colour indicates control coverage strength across prevent, detect, respond, and recover functions.

ATLAS Tactic	Prevent	Detect	Respond	Recover	Pillar	Primary Control
ML Model Access (TA0000)	▮	▮	▮	▮	III	Rate limit + auth
Reconnaissance (TA0001)	▮	▮	▮	▮	II	Registry ACL
Resource Dev. (TA0002)	▮	▮	▮	▮	IV	Red-team prog.
Initial Access (TA0004)	▮	▮	▮	▮	II	ML-BOM verify
Execution (TA0005)	▮	▮	▮	▮	III	SPG + sandbox
Persistence (TA0006)	▮	▮	▮	▮	II	Pipeline integrity
Evasion (TA0007)	▮	▮	▮	▮	II	Robustness test
Impact (TA0008)	▮	▮	▮	▮	III	Kill-switches
Exfiltration (TA0009)	▮	▮	▮	▮	IV	Diff. privacy

Legend: ▮ Strong coverage ▮ Partial coverage ▮ Gap (action required)

B.3 Audit Evidence Requirements

For each ATLAS tactic, governance requires: (a) documented control design rationale; (b) evidence of control testing within the last 90 days; (c) incident correlation showing which controls activated during AI security events; (d) gap analysis identifying techniques with no current mitigation. This matrix forms the core of the AI security evidence pack for regulatory audits and procurement due diligence.

© 2026 Kieran Upadrasta | CISSP, CISM, CRISC, CCSP, MBA, BEng

Appendix C: AI Bill of Materials — SBOM/ML-BOM Artifact Structure

This appendix specifies the structure and content requirements for AI Bills of Materials, covering software (SBOM), machine learning (ML-BOM), and cryptographic (CBOM) components. The specification aligns with CycloneDX 1.6, SPDX 3.0 AI profiles, and the OWASP AIBOM Generator.

C.1 ML-BOM Mandatory Fields

Field	Description	Validation Rule
model.name	Canonical model identifier	Must match model registry entry
model.version	Semantic version of deployed model	Immutable post-deployment; hash-linked
model.type	Architecture type (transformer, CNN, etc.)	Must reference published architecture
model.hash	SHA-256 of model weights file	Verified at deployment and runtime
training.dataset.name	Training dataset identifier	Linked to dataset registry entry
training.dataset.hash	SHA-256 of training data	Immutable; versioned
training.dataset.licence	Data licence/usage rights	Must permit intended use case
training.framework	ML framework and version	e.g., PyTorch 2.4.1, TensorFlow 2.17
training.hardware	Training compute environment	GPU type, cloud region, TEE status
dependencies[]	All Python/system packages	Full SBOM with CVE cross-reference
crypto.algorithms[]	Encryption used for data/weights	CBOM entry; PQC readiness flag
evaluation.metrics	Performance benchmarks at release	Accuracy, F1, bias metrics, adversarial robustness
provenance.attestation	Cryptographic build provenance	SLSA Level 3+ attestation chain

C.2 Example ML-BOM Fragment (CycloneDX 1.6 JSON)

The following illustrates the minimum viable ML-BOM entry for a production credit-scoring model. In practice, the full BOM would include the complete dependency tree, evaluation dataset details, and CBOM entries for all cryptographic operations.

© 2026 Kieran Upadrasta | CISSP, CISM, CRISC, CCSP, MBA, BEng

```
{
  "bomFormat": "CycloneDX",
  "specVersion": "1.6",
  "serialNumber": "urn:uuid:3e671687-395b-41f5-a30f-a58921a69b79",
  "version": 1,
  "components": [{
    "type": "machine-learning-model",
    "name": "credit-risk-scorer-v3",
    "version": "3.2.1",
    "hashes": [{"alg": "SHA-256", "content": "a1b2c3d4..."}],
    "modelCard": {
      "modelParameters": {
        "approach": {"type": "supervised"},
        "task": "binary-classification",
        "architectureFamily": "gradient-boosted-trees",
        "datasets": [{
          "ref": "training-data-credit-2024-q4",
          "type": "training",
          "governance": {"license": {"id": "LicenseRef-Internal-v2"}}
        }]
      },
      "quantitativeAnalysis": {
        "performanceMetrics": [
          {"type": "AUC-ROC", "value": "0.934"},
          {"type": "demographic-parity-gap", "value": "0.02"}
        ]
      }
    }
  }]
}
```

C.3 CBOM Requirements for Post-Quantum Readiness

Every AI system’s BOM must include a Cryptographic Bill of Materials (CBOM) documenting all cryptographic algorithms used for data encryption, model weight protection, attestation signing, and secure communications. The CBOM must flag each algorithm’s post-quantum vulnerability status: (a) PQC-Ready—uses NIST FIPS 203/204/205 compliant algorithms; (b) Migration-Planned—classical algorithm with documented migration timeline to PQC; (c) At-Risk—classical algorithm with no migration plan. The doctrine mandates all new AI systems achieve PQC-Ready status by Q4 2027, with existing systems migrated by 2030 deprecation deadline.

© 2026 Kieran Upadrasta | CISSP, CISM, CRISC, CCSP, MBA, BEng

Appendix D: AI Governance Control Catalogue

This appendix provides a numbered control library with unique identifiers, mapped to ISO 42001 clauses, NIST AI RMF subcategories, EU AI Act articles, and international framework equivalents. Each control includes a maturity scoring rubric enabling self-assessment, audit preparation, and gap analysis.

D.1 Control Numbering Convention

Controls follow the pattern AICD-[DOMAIN]-[NNN] where AICD denotes AI Cyber Doctrine and domains correspond to the four-pillar model:

Domain Code	Pillar	Scope	Control Range
AICD-GOV	I. Governance	Board oversight, policy, roles, risk appetite	AICD-GOV-001 to 010
AICD-SEC	II. Secure by Design	Threat modelling, SDLC, privacy, supply chain	AICD-SEC-001 to 012
AICD-OPS	III. Command & Control	Autonomy levels, kill-switches, monitoring	AICD-OPS-001 to 010
AICD-ACC	IV. Accountability	Logging, lineage, evidence packs, incident	AICD-ACC-001 to 008

D.2 Maturity Scoring Rubric

Each control is assessed on a 0–5 maturity scale aligned with CMMI and ISO 42001 management system requirements:

Level	Label	Description	Evidence Required
0	**Non-existent**	No awareness or implementation of the control	N/A — gap identified
1	**Ad Hoc**	Informal, reactive; dependent on individuals	Interview notes; informal documentation
2	**Developing**	Documented but inconsistently applied	Approved policy; partial implementation records
3	**Defined**	Standardised, documented, communicated organisation-wide	Procedures, training records, role assignments
4	**Managed**	Measured, monitored with KPIs; corrective action taken	Metrics dashboards, audit reports, corrective actions
5	**Optimized**	Continuous improvement;	Trend analysis, automation logs,

© 2026 Kieran Upadrasta | CISSP, CISM, CRISC, CCSP, MBA, BEng

		automated where possible	benchmarking

Target state: All Pillar I (Governance) controls at Level 4+ within 12 months. All Pillar II–IV controls at Level 3+ within 12 months, Level 4+ within 24 months.

© 2026 Kieran Upadrasta | CISSP, CISM, CRISC, CCSP, MBA, BEng

D.3 Pillar I: Governance Controls

Control ID	Control Description	ISO 42001	NIST AI RMF	EU AI Act
AICD-GOV-001	AI Governance Policy: Establish and maintain board-approved AI governance policy	5.2, A.2	GOVERN 1.1	Art. 9(1)
AICD-GOV-002	Board AI Oversight: Designated board committee for AI risk oversight	5.1, A.1	GOVERN 1.2	Art. 9(2)
AICD-GOV-003	AI Risk Appetite: Define risk appetite for AI-driven decisions	A.3, 6.1	GOVERN 1.3	Art. 9(2)(a)
AICD-GOV-004	AI System Inventory: Maintain comprehensive registry of all AI systems	A.5, 7.5	MAP 1.1	Art. 49(1)
AICD-GOV-005	AI Risk Classification: Classify each AI system by EU AI Act risk tier	A.4, 6.1.2	MAP 1.5	Art. 6, Annex III
AICD-GOV-006	RACI Matrix: Defined roles for AI governance, development, operations	5.3, A.2	GOVERN 2.1	Art. 16(a)
AICD-GOV-007	Stakeholder Engagement: Identify and engage affected stakeholders	4.2, A.7	GOVERN 3.1	Art. 9(5)
AICD-GOV-008	Training & Competence: AI-specific training for board, technical, legal	7.2, 7.3	GOVERN 4.1	Art. 4
AICD-GOV-009	Shadow AI Policy: Detection, classification, sanctioning of unsanctioned AI	A.5, 8.1	GOVERN 5.1	Art. 26(1)
AICD-GOV-010	Quarterly Risk Review: Board-level AI risk reporting with FAIR-AIR metrics	9.1, 9.3	MEASURE 1.1	Art. 9(3)

D.4 Pillar II: Secure by Design Controls

Control ID	Control Description	ISO 42001	NIST AI RMF	Intl Mapping
AICD-SEC-001	AI Threat Modelling: STRIDE + ATLAS threat analysis for every AI system	A.6, 8.1	MAP 3.1	UK: Safety; SG: FEAT
AICD-SEC-002	Data Provenance: Cryptographic attestation of training data lineage	A.8, A.9	MAP 2.3	EO 14110 §4.2
AICD-SEC-003	Adversarial Robustness Testing: PGD/AutoAttack on all production models	A.6	MEASURE 2.6	CISA Red-Team
AICD-SEC-	Bias & Fairness Audit:	A.10	MEASURE 2.11	SG: FEAT Fairness

004	Demographic parity, equalised odds assessment			
AICD-SEC-005	Secure Prompt Gateway: Centralised LLM interaction policy enforcement	A.6, 8.1	MANAGE 2.1	CISA SbD
AICD-SEC-006	ML-BOM Generation: Automated BOM creation for all AI components	A.8, 7.5	MAP 2.1	EO 14110 §4.3
AICD-SEC-007	Privacy by Design: DPIA for all AI processing personal data	A.10, A.9	MAP 5.1	UK: Fairness
AICD-SEC-008	Supply Chain Verification: Hash validation of all model artefacts	A.8	MAP 2.2	CISA Supply Chain
AICD-SEC-009	Crypto-Agility: PQC readiness assessment; migration plans for all AI crypto	A.6	MANAGE 4.1	NIST PQC FIPS
AICD-SEC-010	Model Validation: Independent validation of high-risk models (SR 11-7)	A.6, 9.2	MEASURE 2.5	BIS FSI 56

© 2026 Kieran Upadrasta | CISSP, CISM, CRISC, CCSP, MBA, BEng

D.5 Pillar III: Operational Command & Control Controls

Control ID	Control Description	ISO 42001	NIST AI RMF	Intl Mapping
AICD-OPS-001	Autonomy Classification: Define and enforce autonomy levels for each AI system	A.5, 8.1	MANAGE 1.1	UK: Contestability
AICD-OPS-002	Human-in-the-Loop Gates: Mandatory human approval for high-impact decisions	A.5	MANAGE 2.2	Art. 14 (Human)
AICD-OPS-003	Kill-Switch Architecture: Emergency stop with graceful degradation	A.5, 8.1	MANAGE 3.1	Art. 9(4)(e)
AICD-OPS-004	Real-Time Monitoring: Data drift, model performance, behavioural anomalies	9.1, A.6	MEASURE 3.2	SG: AI Verify
AICD-OPS-005	Rate Limiting & Privilege Isolation: Bound AI system actions; least privilege	A.6	MANAGE 2.3	CISA SbD
AICD-OPS-006	Automated Rollback: Revert to last known good model on anomaly detection	A.5, 8.1	MANAGE 4.2	DORA Art. 11
AICD-OPS-007	AI Agent Identity: Unique credential per AI agent; machine-to-machine auth	A.6	MANAGE 2.4	NIST SP 800-207
AICD-OPS-008	Shadow AI Detection: CASB integration, outbound API monitoring	A.5, 9.1	GOVERN 5.2	Art. 26(1)

D.6 Pillar IV: Accountability Controls

Control ID	Control Description	ISO 42001	NIST AI RMF	Intl Mapping
AICD-ACC-001	Decision Audit Trail: Complete lineage for every algorithmic decision	A.8, 7.5	MANAGE 1.3	UK: Transparency
AICD-ACC-002	Model Cards: Standardised documentation for all production models	A.8, 7.5	MAP 1.3	Art. 11 (Tech Doc)
AICD-ACC-003	Explainability: Interpretable outputs for high-risk/regulated decisions	A.10	MEASURE 2.8	SG: FEAT Explain.
AICD-ACC-004	Incident Response (AI): ATLAS-classified IR playbook with 72-hour NIS2 notification	A.7, 10.1	MANAGE 4.1	NIS2 Art. 23
AICD-ACC-005	Evidence Pack Generation: Automated regulatory evidence for audit readiness	9.2, 7.5	GOVERN 6.1	Art. 12 (Records)

© 2026 Kieran Upadrasta | CISSP, CISM, CRISC, CCSP, MBA, BEng

AICD-ACC-006	Post-Market Monitoring: Continuous performance surveillance for deployed AI	A.6, 9.1	MANAGE 3.2	Art. 72 (Post-Mkt)
AICD-ACC-007	Tabletop Exercises: Quarterly AI-specific incident simulation	A.7, 10.1	MANAGE 4.3	DORA Art. 26
AICD-ACC-008	Third-Party AI Assurance: Independent audit of high-risk AI systems	9.2, 9.3	GOVERN 6.2	Art. 43 (Conform.)

D.7 Cross-Reference Summary: Framework Alignment

The following summarises how the 36 AICD controls map across the primary international frameworks, enabling a single implementation to satisfy multiple regulatory obligations simultaneously:

Framework	Controls Mapped	GOV	SEC	OPS	ACC	Coverage
ISO 42001:2023	**36/36**	10	10	8	8	**100%**
NIST AI RMF 1.0	**36/36**	10	10	8	8	**100%**
EU AI Act	**34/36**	10	10	7	7	**94%**
UK AI Principles	**28/36**	6	8	6	8	**78%**
Singapore FEAT	**22/36**	4	8	4	6	**61%**
CISA AI Roadmap	**20/36**	2	10	6	2	**56%**
US EO 14110	**18/36**	4	8	4	2	**50%**

> *A single implementation of these 36 controls achieves 100% coverage of ISO 42001 and NIST AI RMF, 94% of EU AI Act requirements, and substantial coverage of UK, Singapore, and US obligations. This cross-reference mapping enables multi-jurisdictional compliance through a unified governance baseline.*

© 2026 Kieran Upadrasta | CISSP, CISM, CRISC, CCSP, MBA, BEng

Appendix E: Conformity Assessment Methodology

This appendix defines the conformity assessment methodology for independent evaluation of an organisation's implementation of this framework. The methodology aligns with ISO/IEC 17021-1:2015 (requirements for certification bodies) and ISO 19011:2018 (auditing management systems).

E.1 Conformity Scoring Aggregation Model

Each of the 36 controls (AICD-GOV-001 through AICD-ACC-008) is assessed using the 0–5 maturity scale defined in Appendix D.2. Conformity is determined through a three-tier aggregation:

Assessment Level	Calculation	Pass Threshold	Scope
Individual Control	Maturity score (0–5)	≥3 (Defined)	Each AICD control
Pillar Score	Weighted mean of controls in pillar	≥3.0 pillar average	GOV, SEC, OPS, ACC
Framework Score	Weighted mean of all 4 pillars	≥3.5 overall	Full framework

Pillar weighting reflects regulatory emphasis:

Pillar	Weight	Controls	Max Score	Pass
I. Governance	30%	10	50	**≥30**
II. Secure by Design	25%	10	50	**≥30**
III. Command & Control	25%	8	40	**≥24**
IV. Accountability	20%	8	40	**≥24**

Conformity determination: an organisation achieves conformity when (a) no individual control scores below 2 (Developing); (b) each pillar average reaches ≥3.0; (c) overall framework score reaches ≥3.5; and (d) no major nonconformities remain open.

E.2 Audit Sampling Methodology

The audit sampling approach determines the evidence depth required per control, scaled to organisational complexity and AI system portfolio size.

Organisation Tier	AI Systems	Minimum	Evidence Depth	Audit Duration

© 2026 Kieran Upadrasta | CISSP, CISM, CRISC, CCSP, MBA, BEng

		Sample		
Tier 1: Large Enterprise	50+ AI systems	√(n) + 2	Full evidence pack per sampled system	10–15 audit days
Tier 2: Mid-Market	10–49 AI systems	√(n) + 1	Targeted evidence for high-risk; sample for others	6–10 audit days
Tier 3: SME / Startup	1–9 AI systems	All systems	Proportionate evidence	3–5 audit days

For Tier 1 organisations with 100 AI systems, minimum sample size = √(100) + 2 = 12 systems. All high-risk AI systems (EU AI Act Annex III) must be sampled regardless of portfolio size. Sampling must cover at least one system per business unit deploying AI.

E.3 Nonconformity Classification Taxonomy

Nonconformities identified during assessment are classified using a three-tier severity model:

Severity	Definition	Remediation	Impact on Conformity
MAJOR	Absence of or total failure of a control, or a systematic failure pattern across multiple controls	Corrective action plan within 30 days; evidence of closure within 90 days	**Conformity withheld until closed**
MINOR	Partial implementation, inconsistent application, or incomplete evidence for a specific control	Corrective action within 60 days; verified at next surveillance	Conformity granted with conditions
OFI	Opportunity for improvement: control is conformant but optimisation potential identified	Recommendation noted; no mandatory action	No impact

Escalation rules: (a) three or more minor nonconformities within a single pillar constitute a major nonconformity for that pillar; (b) any control scoring 0 (Non-existent) for a high-risk AI system is automatically classified as major; (c) repeat minor nonconformities from prior audit that remain unresolved escalate to major.

E.4 Surveillance and Recertification Cycle

Audit Type	Frequency	Scope	Output
Initial Certification	Year 0	Full assessment of all 36 controls	Conformity determination + score
Surveillance 1	Year 0 + 12 months	50% of controls (rotated); all open NCs; any new AI systems	Surveillance report; NC status

© 2026 Kieran Upadrasta | CISSP, CISM, CRISC, CCSP, MBA, BEng

Surveillance 2	Year 0 + 24 months	Remaining 50% of controls; all open NCs; material changes	Surveillance report; NC status
Recertification	Year 0 + 36 months	Full re-assessment of all 36 controls	Recertification decision
Triggered Audit	As needed	Triggered by: material AI incident, regulatory finding, significant scope change, or organisational restructuring	Special assessment report

The three-year cycle aligns with ISO management system certification norms. Surveillance audits sample approximately 50% of controls on rotation, ensuring full coverage across the cycle. High-risk AI systems are assessed at every audit event.

© 2026 Kieran Upadrasta | CISSP, CISM, CRISC, CCSP, MBA, BEng

Appendix F: Reproducible FAIR-AIR Risk Quantification Dataset

This appendix provides calibrated reference data for FAIR-AIR risk quantification, enabling organisations to benchmark their own AI risk scenarios against industry baselines. The dataset draws on published sources and is designed for reproducible application across sectors.

F.1 Threat Event Frequency Calibration

Base rates for AI-specific threat events, derived from industry incident reporting:

Threat Event	Low (10th %ile)	Likely (50th %ile)	High (90th %ile)	Very High (95th %ile)	Source Basis
Prompt injection attempt	50/yr	200/yr	800/yr	2,000/yr	OWASP LLM 2025
Adversarial model manipulation	2/yr	12/yr	40/yr	80/yr	MITRE ATLAS cases
Data poisoning (training)	0.5/yr	3/yr	15/yr	30/yr	JFrog 2025 report
Shadow AI security incident	5/yr	24/yr	60/yr	120/yr	IBM 2025 breach
Deepfake-enabled fraud	1/yr	6/yr	20/yr	50/yr	Deloitte 2025
Agentic AI misuse / exploit	1/yr	8/yr	30/yr	60/yr	Gartner 2025
Model drift causing compliance breach	2/yr	10/yr	25/yr	50/yr	SR 11-7 guidance

F.2 Loss Magnitude Calibration (€, per event)

Single Loss Expectancy ranges by AI incident type and organisational scale:

Loss Scenario	Minimum	Most Likely	Maximum	Distribution
Regulatory fine (EU AI Act)	€100K	€2.5M	€35M	Log-normal
Data breach (AI-involved)	€500K	€4.44M	€15M	Log-normal
D&O litigation (AI-related)	€2M	€52M	€200M	Power-law tail
Fraud loss (deepfake-enabled)	€50K	€2M	€25M	Log-normal
Operational disruption (model failure)	€200K	€1.5M	€10M	PERT
Reputational damage (algorithmic bias)	€500K	€5M	€50M	Triangular

© 2026 Kieran Upadrasta | CISSP, CISM, CRISC, CCSP, MBA, BEng

F.3 Control Effectiveness Reduction Factors

The following table provides empirically calibrated reduction factors for each control domain when implemented at maturity Level 3 or above. These factors are applied to the FAIR Vulnerability component:

Control Domain	No Control (Level 0)	Ad Hoc (Level 1)	Defined (Level 3)	Optimized (Level 5)	Reduction Range
AI Governance Policy (GOV-001)	1.00	0.85	0.55	0.30	15–70%
Secure Prompt Gateway (SEC-005)	1.00	0.75	0.30	0.10	25–90%
Kill-Switch Architecture (OPS-003)	1.00	0.80	0.40	0.15	20–85%
Decision Audit Trail (ACC-001)	1.00	0.90	0.60	0.35	10–65%
ML-BOM + Supply Chain (SEC-006/008)	1.00	0.80	0.45	0.20	20–80%
Shadow AI Detection (OPS-008)	1.00	0.85	0.50	0.25	15–75%

Interpretation: a Vulnerability factor of 0.30 at Level 3 means the control reduces the probability of threat event success by 70%. These factors are intended as calibration starting points; organisations should refine them using their own incident data through Bayesian updating over successive measurement periods.

F.4 Worked Example: Annualised Loss Expectancy Calculation

The following demonstrates a complete FAIR-AIR calculation using reference data from this appendix:

FAIR-AIR Component	Value	Source	Notes
Scenario	Adversarial model manipulation	—	Credit model
Threat Event Frequency (TEF)	12/year (50th %ile)	Table F.1	Baseline
Vulnerability (pre-control)	0.72	Expert estimate	No SPG, no hardening
SPG Reduction (Level 3)	0.30 factor	Table F.3	70% reduction
Vulnerability (post-control)	0.72 × 0.30 = 0.22	Calculated	
Loss Event Frequency (LEF)	**12 × 0.22 = 2.6/year**	TEF × Vuln	

© 2026 Kieran Upadrasta | CISSP, CISM, CRISC, CCSP, MBA, BEng

Single Loss Expectancy (SLE)	€2.5M (most likely)	Table F.2	Regulatory fine
Annualised Loss Expectancy (ALE)	**2.6 × €2.5M = €6.5M**	LEF × SLE	Post-SPG
ALE without controls	**12 × 0.72 × €2.5M = €21.6M**	Baseline	No SPG
Risk Reduction	**€21.6M − €6.5M = €15.1M**	Delta	From one control

This worked example demonstrates the calculation methodology using a single control. In practice, organisations should apply Monte Carlo simulation across all relevant threat scenarios, aggregating control effectiveness factors multiplicatively where controls address different attack stages, and taking the maximum reduction where controls overlap on the same attack stage.

Reproducibility note: All input parameters in Tables F.1–F.3 are documented with source references. Organisations using this dataset should record any calibration adjustments, the rationale for those adjustments, and the resulting outputs as part of the audit evidence required by Appendix E. This enables third-party auditors to independently verify risk quantification calculations.

© 2026 Kieran Upadrasta | CISSP, CISM, CRISC, CCSP, MBA, BEng

About the Author

Kieran Upadrasta

CISSP | CISM | CRISC | CCSP | MBA | BEng

Kieran Upadrasta brings 27 years of cybersecurity experience across all Big 4 firms (Deloitte, PwC, EY, KPMG) and 21 years in financial services and banking. He holds CISSP, CISM, CRISC, CCSP certifications alongside an MBA and BEng, and serves as Professor of Practice in Cybersecurity, AI, and Quantum Computing. He leads the PRMIA Cyber Security Programme and serves as Lead Auditor at ISF Auditors and Control. His track record spans 40+ enterprise transformations, 12+ regulatory jurisdictions, and €500B+ in aggregate asset environments governed.

BRE-14142026 | Version 7.0 | March 2026 | All Rights Reserved

www.kie.ie | info@kieranupadrasta.com

References

[1] IBM, Cost of a Data Breach Report 2025, IBM Security, July 2025.

[2] SecurityWeek, "AI-Assisted Cyberattacks Surge 72% Year-Over-Year," SecurityWeek, March 2025.

[3] European Parliament, Regulation (EU) 2024/1689 of 13 June 2024 (EU AI Act), Official Journal of the European Union, 2024.

[4] Techne Analytics, "D&O Settlements in AI-Related Securities Class Actions," Techne AI Liability Report, Q1 2025.

[5] European Parliament, Regulation (EU) 2022/2554 (DORA), Official Journal of the European Union, January 2023.

[6] European Parliament, Directive (EU) 2022/2555 (NIS2 Directive), Official Journal of the European Union, December 2022.

[7] US Securities and Exchange Commission, "SEC Creates Cyber and Emerging Technologies Unit," Press Release, February 2025.

[8] NIST, Artificial Intelligence Risk Management Framework (AI RMF 1.0), NIST, January 2023.

[9] ISO/IEC 42001:2023, Information Technology — Artificial Intelligence — Management System, International Organization for Standardization, 2023.

[10] Gartner, "Organizations Aligning NIST AI RMF and ISO 42001 Achieve 3.4x Higher Governance Effectiveness," Gartner Research Note, Q2 2025.

[11] OWASP, "OWASP Top 10 for LLM Applications 2025," OWASP Foundation, 2025.

[12] Deloitte, "Generative AI-Facilitated Fraud Losses Projections 2023–2027," Deloitte Center for Financial Services, 2025.

[13] Gartner, "Top Strategic Technology Trends for 2026: Agentic AI Security," Gartner, October 2025.

[14] MITRE, "MITRE ATLAS: Adversarial Threat Landscape for AI Systems," MITRE Corporation, 2025.

[15] FAIR Institute, "FAIR-AIR Playbook: Quantifying AI Risk," FAIR Institute, 2025.

[16] JFrog, "2025 Software Supply Chain State of the Union Report," JFrog, February 2025.

[17] CycloneDX, "CycloneDX 1.6 ML-BOM Support," OWASP CycloneDX, 2024.

[18] NIST, "NIST Releases Three Post-Quantum Cryptography Standards (FIPS 203, 204, 205)," NIST, August 2024.

[19] Forrester, "Quantum Security Spending Projections 2026," Forrester Research, Q4 2025.

[20] Delaware Court of Chancery, In re Caremark International Inc. Derivative Litigation, 698 A.2d 959 (Del. Ch. 1996).

[21] National Association of Corporate Directors (NACD), "Board AI Governance Framework," NACD, 2025.

[22] CISA, "Secure by Design Principles for AI Systems," Cybersecurity and Infrastructure Security Agency, 2025.

[23] NIST SP 800-207, "Zero Trust Architecture," NIST Special Publication, August 2020.

[24] Gartner, "IAM Leaders Must Uniquely Identify Each AI Agent and Enforce Least Privilege," Gartner IAM Guidance, 2025.

[25] OWASP, "AIBOM Generator: First Open-Source Tool for AI Software Bills of Materials," OWASP Foundation, 2025.

[26] SR 11-7, "Guidance on Model Risk Management," Board of Governors of the Federal Reserve System / OCC, April 2011.

[27] European Commission, "DigitalJustice@2030 Strategy," European Commission, November 2025.

[28] UK HMCTS, "Reform Programme Annual Report 2024–25," HM Courts and Tribunals Service, 2025.

[29] Gartner, "40%+ of Agentic AI Projects Will Be Cancelled by End of 2027," Gartner Press Release, 2025.

[30] SPDX, "SPDX 3.0 Specification with AI and Dataset Profiles," Linux Foundation, 2024.

[31] The Open Group, "Factor Analysis of Information Risk (FAIR) Standard, Version 4," The Open Group, 2024.

[32] Ponemon Institute, "The Cost of AI Security Incidents in Financial Services," Ponemon/IBM, Q3 2025.

[33] Executive Office of the President, "Executive Order 14110: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence," Federal Register, October 30 2023.

[34] Infocomm Media Development Authority (IMDA), "Model AI Governance Framework, Second Edition," Singapore, January 2020; updated with AI Verify Foundation, 2024.

© 2026 Kieran Upadrasta | CISSP, CISM, CRISC, CCSP, MBA, BEng

[35] UK Department for Science, Innovation and Technology, "A Pro-Innovation Approach to AI Regulation," HM Government White Paper, August 2023; updated March 2025.

[36] CISA, "Roadmap for Artificial Intelligence," Cybersecurity and Infrastructure Security Agency, November 2023; updated 2025.

[37] Bank for International Settlements, "Artificial Intelligence in Financial Services: Governance and Regulatory Considerations," BIS FSI Insights No 56, March 2024.

[38] Monetary Authority of Singapore, "Principles to Promote Fairness, Ethics, Accountability and Transparency (FEAT) in the Use of AI and Data Analytics," MAS, 2024.

[39] ISO/IEC 17021-1:2015, "Conformity Assessment — Requirements for Bodies Providing Audit and Certification of Management Systems," International Organization for Standardization, 2015.

[40] ISO 19011:2018, "Guidelines for Auditing Management Systems," International Organization for Standardization, 2018.

[41] Hubbard, D.W. and Seiersen, R., "How to Measure Anything in Cybersecurity Risk," Wiley, 2nd edition, 2023.

[42] Freund, J. and Jones, J., "Measuring and Managing Information Risk: A FAIR Approach," Butterworth-Heinemann, 2nd edition, 2024.

© 2026 Kieran Upadrasta | CISSP, CISM, CRISC, CCSP, MBA, BEng