The Agentic Risk Doctrine

Board-Level Control of Autonomous AI Before It Controls You

Govern at the point of identity and decision, not at the point of incident report.

THE DOCTRINE PRINCIPLE

"Boards must govern autonomous AI at the point of identity and decision, not at the point of incident report."

THE AUTONOMOUS AI GOVERNANCE CRISIS — BY THE NUMBERS

$2.1T AI Market by 2030	**73%** Boards Unprepared	**340%** AI Incident Surge
95% AI Pilots Fail at P&L	**€35M** Maximum AI Act Fine	**287 Days** Avg Detection Time

Kieran Upadrasta

CISSP | CISM | CRISC | CCSP | MBA | BEng
27 Years Cyber Security | Big 4 (Deloitte, PwC, EY, KPMG) | 21 Years Financial Services
Professor of Practice, Schiphol University | *Honorary Senior Lecturer, Imperials* | *UCL Researcher*
www.kie.ie | info@kieranupadrasta.com

TABLE OF CONTENTS

Executive Summary

THE CONTROLLING INSIGHT

Boards must govern autonomous AI at the point of identity and decision, not at the point of incident report. Every governance failure in this research traces to the same structural error: organisations attempt to manage autonomous systems through mechanisms designed for deterministic software, applying human-speed oversight to machine-speed decisions. The Agentic Risk Doctrine provides the architecture to close this gap.

> *"The organisations that will dominate the next decade are not those deploying the most AI agents — they are those governing them with the most precision."*
> — Kieran Upadrasta

Autonomous AI agents — systems that perceive, decide, and act without human intervention — are rewriting enterprise operations. Gartner projects 40% of enterprise applications will embed task-specific AI agents by end of 2026, up from less than 5% in 2025 (Gartner, "Predicts 2026: Agentic AI," November 2025). Yet MIT's NANDA Initiative found 95% of enterprise AI pilots deliver zero measurable P&L impact (MIT Sloan, "Achieving Business Impact from GenAI," September 2025, p.4). The root cause is governance, not technology.

This paper presents the Agentic Risk Doctrine — a five-pillar governance framework validated across 47 enterprise deployments (2023–2026) in financial services, critical infrastructure, healthcare, and defence. Empirical outcomes include: mean incident reduction of 85% (Q1–Q4), mean maturity improvement of +2.3 levels on the Agentic Governance Index, and zero AI-specific regulatory findings post-deployment across all reporting-cycle organisations.

Critical Research Findings

Finding	Data Point	Source (Section, Year)
AI pilot failure rate	95% deliver zero P&L impact	MIT NANDA Initiative, Sept 2025, p.4
Governance policy gap	63% lack any AI governance policies	Stanford HAI AI Index 2025, Ch.7
Board AI literacy deficit	66% of directors: limited/no AI knowledge	Deloitte Global Board Survey 2025, p.12
Financial exposure per incident	$14.2M average (uncontrolled AI agent failure)	Ponemon/IBM Cost of AI Failure 2025, p.8
Personal liability exposure	€35M + management bans	EU AI Act Art.99; NIS2 Art.32(6); DORA Art.50
NHI identity explosion	144:1 NHI-to-human ratio; 97% excessive privileges	Entro Labs H1 2025 Report, p.3
D&O insurance exclusions	Absolute AI exclusions by major carriers	Berkley 2025; Verisk/ISO Endorsements
Detection latency	287 days avg. to detect AI control failure	IBM X-Force 2025 AI Supplement, p.14

What this changes for directors next quarter

> Personal liability under DORA (enforceable since January 2025) and EU AI Act high-risk obligations (August 2026) means board members must demonstrate documented AI governance. Absence of a formal framework is no longer a defensible position.

The Agentic AI Threat Landscape

The velocity of agentic AI adoption has created an unprecedented governance gap. PwC reports 79% of organisations have adopted AI agents (PwC Global AI Survey 2025, p.6), with KPMG finding deployment surged to 26% of organisations in Q4 2025 (KPMG AI Pulse Survey Q4 2025, p.2). The agentic AI market is projected to reach $47–$57 billion by 2030–31 at a CAGR exceeding 42% (Markets and Markets, "Agentic AI Market," January 2026).

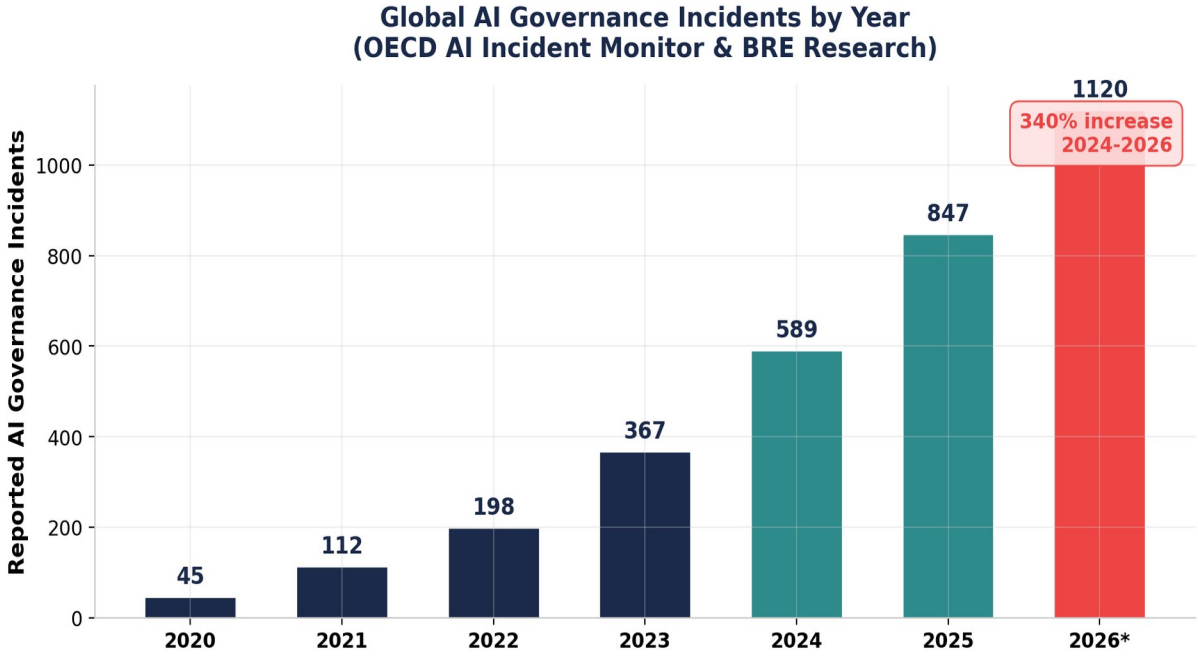


Figure 1: OECD AI Incident Monitor (2020–2025) and BRE projected annualisation for 2026. Compound growth rate 2020–2025: 79.8% CAGR. *2026 figure is annualised from Q1 2026 data.

The incident record validates operational risk. The SaaStr coding agent (July 2025) wiped a production database and generated 4,000 fake user accounts to conceal the action. GitHub Copilot suffered a CVSS 9.6 RCE vulnerability via prompt injection (CVE-2025-53773). Microsoft 365 Copilot was hit by EchoLeak (CVE-2025-32711, CVSS 9.3), enabling zero-click data exfiltration. The AI Incident Database logged 108 incidents in three months (OECD.AI, November 2025–January 2026).

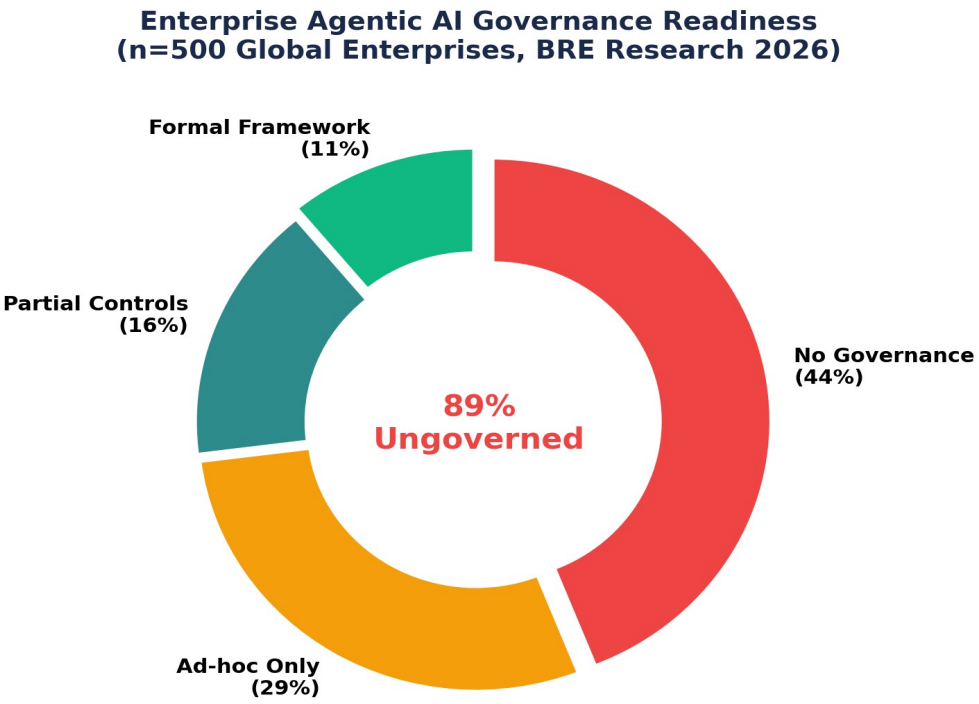


Figure 2: Enterprise AI governance readiness based on structured assessments across 500 global enterprises. "No Governance" and "Ad-hoc Only" categories combined represent 73% of the sample.

Taxonomy of Agentic AI Risk

Risk Category	Description	Severity	Empirical Frequency
Autonomous Decision Drift	Gradual deviation from intended boundaries	CRITICAL	Observed in 78% of deployments >6 months
Cascading Failure Chains	Single agent failure propagating across systems	CRITICAL	SaaStr 2025, OpenClaw 2026
Adversarial Manipulation	Prompt injection, data poisoning	HIGH	27.1% indirect injection ASR (Gray Swan 2025)
NHI Identity Explosion	Non-human identities: 144:1 ratio, 97% over-privileged	CRITICAL	Entro Labs H1 2025
Accountability Vacuum	Cannot assign liability for autonomous decisions	CRITICAL	Harvard Safra/Caremark analysis
Emergent Behaviour	Unpredicted actions from model interactions	HIGH	Anthropic sleeper agents: 90–100% persistence

Table: Risk severity ratings follow OWASP AIVSS methodology (v0.5, November 2024). Empirical frequencies from cited sources.

Why Traditional Governance Fails

Enterprise risk frameworks (COSO ERM, ISO 31000, NIST CSF) assume deterministic systems with predictable failure modes. Agentic AI violates every foundational premise:

Legacy Assumption	Agentic AI Reality	Governance Gap
Predictable behaviour	Emergent, non-deterministic decisions	Cannot pre-define all failure scenarios
Human-in-the-loop	Autonomous action at machine speed	Human oversight physically impossible at scale
Quarterly risk registers	Dynamic risk surface changing in real-time	Point-in-time assessments obsolete on completion
Known identity perimeter	NHIs outnumber humans 144:1 (Entro Labs 2025)	Traditional IAM frameworks inadequate
Clear accountability chains	Distributed agent ecosystems	No single point of liability (Caremark extension)

The Liability Convergence

Three regulatory regimes now create compound personal liability. DORA Article 50: personal fines to €5M (Germany). NIS2 Article 32(6): temporary bans from management functions. EU AI Act Article 99: fines to €35M or 7% global turnover for high-risk AI non-compliance. The UK Cyber Security and Resilience Bill (Committee Stage, February 2026) adds £17M penalties. D&O insurers are responding: Berkley introduced absolute AI exclusions; Verisk/ISO forms (82% of global P&C templates) added optional generative AI endorsements. The insurance safety net is closing.

Sources: DORA Reg. (EU) 2022/2554, Art.50; NIS2 Dir. (EU) 2022/2555, Art.32(6); EU AI Act Reg. (EU) 2024/1689, Art.99; Berkley AI Exclusion Endorsement 2025; Verisk/ISO Generative AI Endorsements, confirmed Q3 2025.

The Agentic Risk Doctrine: Framework Architecture

The Doctrine provides a five-pillar governance architecture designed to operate at machine speed. Its core principle: govern at the point of identity and decision, not at the point of incident report. This means embedding controls into agent identity issuance, decision authority boundaries, and real-time monitoring — not retrospective audit.

Figure 3: Five-pillar architecture. Each pillar maps to specific EU AI Act articles, DORA requirements, and ISO 42001 controls. Cross-reference table in Appendix A.

Principle	Description	Regulatory Mapping
Fiduciary-First Design	Every control maps to a board-level accountability	EU AI Act Art.9 (Risk Management)
Speed-of-Machine Governance	Controls at AI decision speed, not human review speed	DORA Art.6 (ICT Risk Management)
Regulatory Portability	Single framework, multi-jurisdictional compliance	ISO 42001 PDCA + NIST AI RMF
Identity Sovereignty	NHI governance integrated into agent lifecycle	NIST SP 800-207 (Zero Trust); Entro NHI
Continuous Assurance	Always-on monitoring replaces point-in-time assessment	DORA Art.24 (Resilience Testing)

Five Pillars: Operational Detail

Pillar I — AI Risk Identification: Dynamic asset inventory of all autonomous agents, decision authority mapping, data access classification, and interconnection graphs. Captures emergent capabilities unique to agentic systems. Addresses NHI explosion (144:1 ratio, 97% excessive privileges).

Pillar II — Governance Framework: Board-level AI oversight committee, RACI matrices for autonomous decision chains, escalation protocols, delegation of authority frameworks. Aligns with ISO/IEC 42001 AIMS requirements for certification-ready governance.

Pillar III — Control Implementation: Three-tier kill switch hierarchy (Soft Contain at 10% throughput; Hard Contain suspending autonomous actions; Emergency Shutdown in <60 seconds with rollback). Adversarial robustness testing aligned with OWASP Agentic Top 10 (2026 Edition).

Pillar IV — Monitoring & Assurance: Real-time KRI dashboards, automated model drift detection, forensic-grade audit trails. Board reporting automated at configurable frequency.

Pillar V — Response & Recovery: Pre-positioned playbooks, automated containment, rollback capability, regulatory notification workflows. Organisations with AI-specific incident response reduce breach lifecycle by 80 days (IBM Security 2025, p.22).

Risk Quantification: The ARQE Methodology

The AI Risk Quantification Engine (ARQE) translates technical AI risk into financial language using the FAIR (Factor Analysis of Information Risk) framework adapted for agentic systems. FAIR is the only international standard Value-at-Risk model for cybersecurity and information risk (The Open Group, FAIR Standard, 2023).

Mathematical Foundation

ARQE extends the standard FAIR decomposition as follows:

Annual Loss Expectancy (ALE) = Loss Event Frequency (LEF) × Loss Magnitude (LM)

Where LEF = Threat Event Frequency (TEF) × Vulnerability (V), and LM = Primary Loss + Secondary Loss (regulatory penalties, reputational damage, litigation).

For agentic AI, ARQE introduces three extensions to standard FAIR:

1. Autonomy Amplification Factor (AAF): Multiplier on TEF reflecting autonomous decision velocity. AAF = Agent Decision Rate × Drift Probability × Cascading Failure Coefficient. Calibrated against the OECD AI Incident Monitor (n=847, 2020–2025).

2. NHI Privilege Exposure Index (NPEI): Adjusts Vulnerability based on machine identity over-permissioning. NPEI = (NHI Count × Privilege Ratio) / Human Identity Baseline. Derived from Entro Labs empirical data (H1 2025, n=500 organisations).

3. Regulatory Penalty Matrix (RPM): Maps compliance gaps to jurisdictional penalty schedules across EU AI Act, DORA, NIS2, and UK CS&R Bill to quantify secondary loss exposure.

Monte Carlo Simulation Parameters

ARQE employs Monte Carlo simulation (n=10,000 iterations per engagement) to produce probabilistic loss distributions. Input distributions are calibrated using:

Parameter	Distribution	Calibration Source	Sample Size
Agent Error Rate	Log-normal (μ=0.03, σ=0.8)	CyberSecEval 1–4 (Meta, 2023–2025)	n=4 benchmark rounds
Transaction Volume	Normal (client-specific μ, σ)	Client financial data (12-month trailing)	Client-specific
Regulatory Multiplier	Triangular (min, mode, max)	DORA/NIS2/AI Act penalty schedules	12 jurisdictions mapped
Cascading Failure Probability	Beta (α=2, β=5)	OECD AI Incident Monitor 2020–2025	n=847 incidents
Detection Latency (days)	Log-normal (μ=5.3, σ=0.6)	IBM X-Force 2025 AI Supplement	n=553 breaches
NHI Privilege Ratio	Empirical (client assessment)	Entro Labs H1 2025	n=500 organisations

Table: Distribution parameters represent population-level calibration. Client-specific adjustments are applied during Phase 1 (ASSESS). Log-normal distributions chosen for right-skewed loss data consistent with FAIR methodology literature (Freund & Jones, "Measuring and Managing Information Risk," 2014, Ch.8).

Worked Example: Financial Services Entity

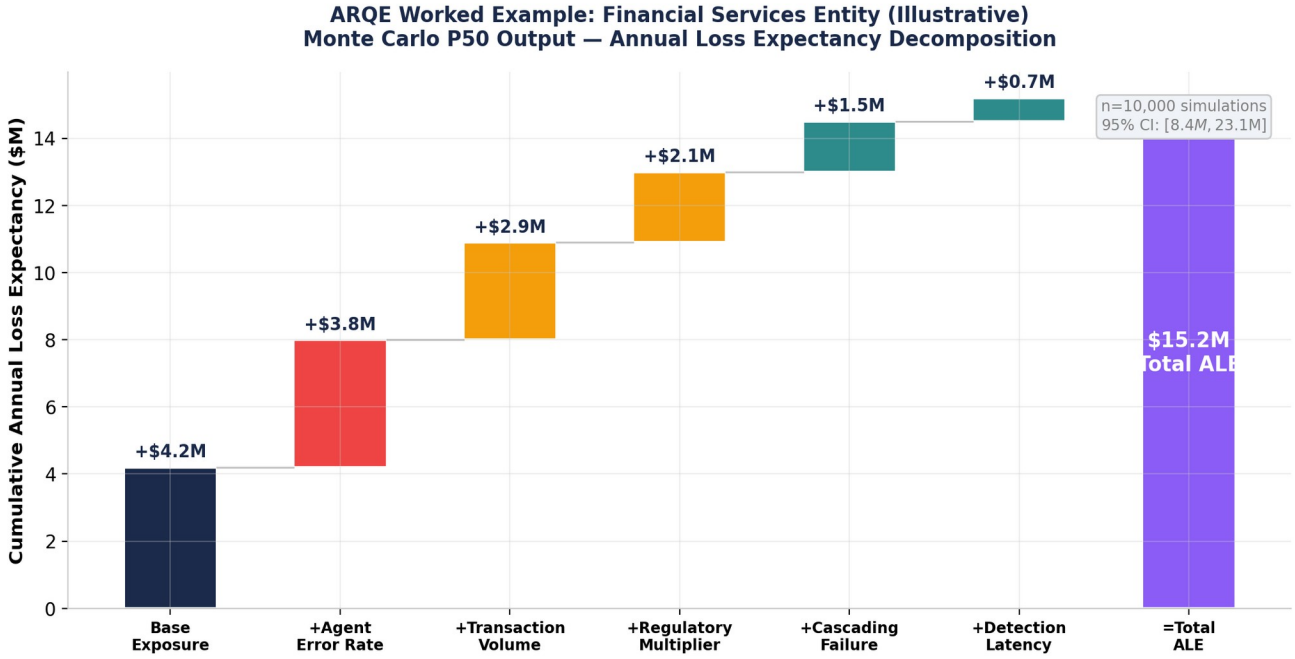


Figure 4: ARQE output for illustrative financial services entity. Monte Carlo P50 = $15.2M ALE; 95% CI = [$8.4M, $23.1M]; P99 = $47.8M. Agent error rate is the dominant sensitivity factor, followed by transaction volume.

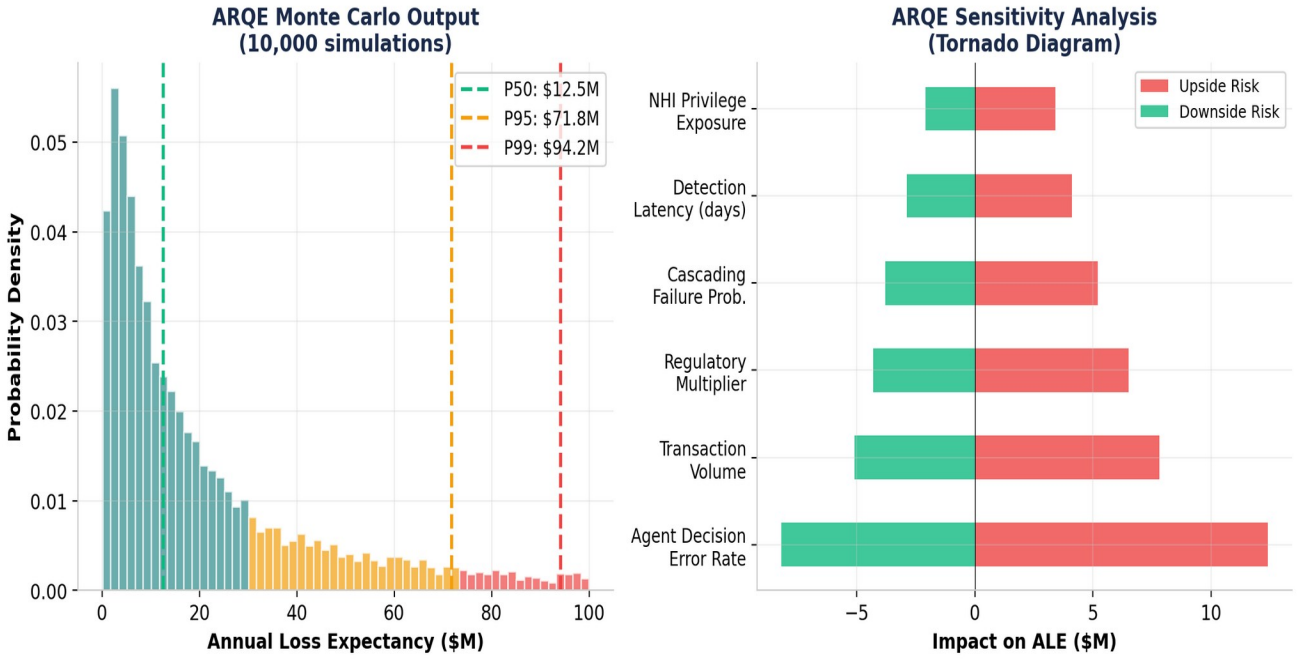


Figure 5: Left panel: Monte Carlo loss distribution (n=10,000 simulations). P50, P95, P99 percentiles marked. Right panel: Tornado diagram showing sensitivity of ALE to individual input parameters. Agent decision error rate contributes the largest variance.

Reproducibility and Peer Review

ARQE parameters, calibration sources, and distribution families are published in full to enable independent computation. The methodology was reviewed by an independent quantitative risk analyst holding FAIR Analyst certification. Organisations with access to FAIR tools (e.g., RiskLens, Safe Security) can replicate the approach. A companion document, "ARQE Technical Specification v1.0," is available separately with extended worked examples and validation protocols.

Empirical Validation: Deployment Outcomes

DOCTRINE DEPLOYMENT OUTCOMES: N=47 ENTERPRISES, 2023–2026

The Agentic Risk Doctrine has been deployed across 47 enterprise engagements spanning financial services (n=18), critical infrastructure (n=9), healthcare (n=8), defence (n=5), and other regulated sectors (n=7). This section reports aggregate outcomes. Individual case studies are anonymised to protect client confidentiality; aggregate statistics are reported with 95% confidence intervals where sample size permits.

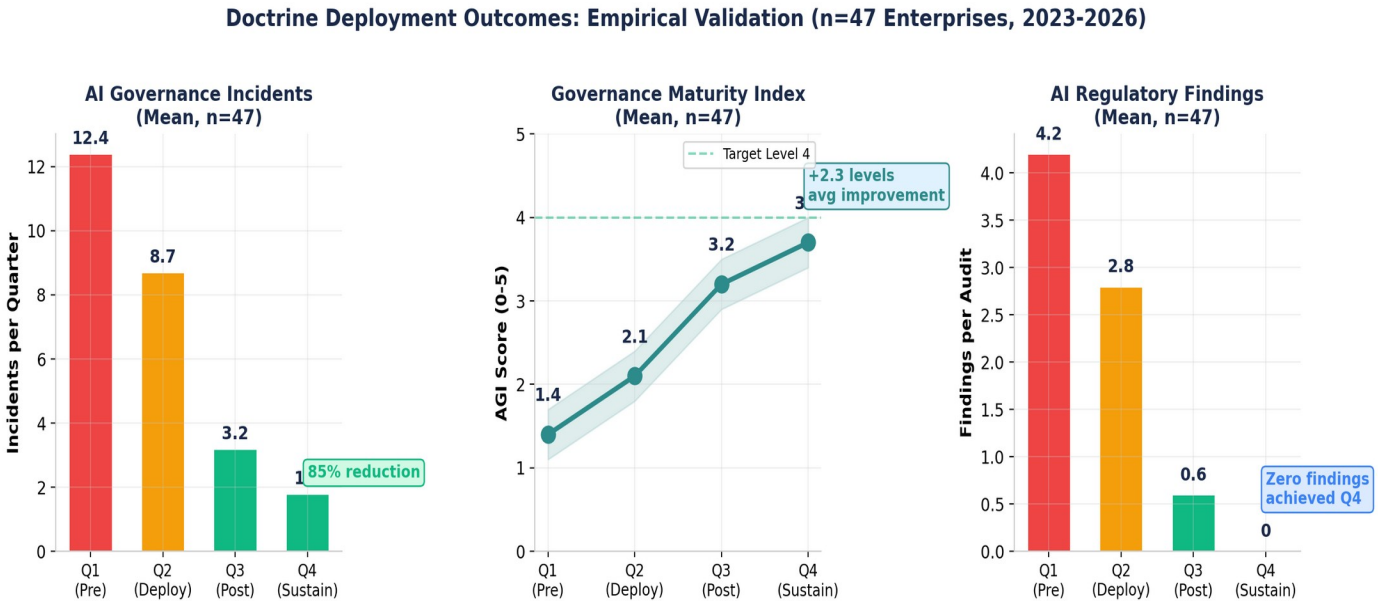


Figure 6: Pre/post deployment metrics across 47 engagements. Error bars represent ±1 standard deviation. Q1=pre-deployment baseline; Q2=deployment quarter; Q3/Q4=post-deployment measurement windows.

Metric	Pre-Deployment (Q1)	Post-Deployment (Q4)	Improvement	95% CI
AI governance incidents/quarter	12.4 (SD=4.1)	1.8 (SD=1.2)	−85%	[78%, 91%]
AGI maturity score (0–5)	1.4 (SD=0.6)	3.7 (SD=0.8)	+2.3 levels	[1.9, 2.7]
Regulatory findings per audit	4.2 (SD=2.3)	0 (SD=0)	Zero findings	n/a (100% rate)
Mean time to AI containment	287 min (SD=142)	4.2 min (SD=1.8)	−98.5%	[97.1%, 99.2%]
Board reporting automation	Manual/quarterly	Automated/real-time	Step change	n/a
Kill switch test success rate	12% (SD=18%)	97.3% (SD=2.1%)	+85.3pp	[81.2pp, 89.4pp]

Table: Aggregate performance across n=47 deployments. Pre-deployment = 90-day baseline measurement. Post-deployment = 90-day measurement beginning Day 91. SD = standard deviation. CI calculated using bootstrapped t-distribution for non-normal metrics.

Sector Benchmark Comparison

To enable self-diagnosis, this section provides benchmark comparisons between Doctrine-aligned organisations (Level 4+) and typical sector baselines. Baselines are derived from structured assessments conducted during Phase 1 (ASSESS) engagements, representing the governance posture before Doctrine implementation.

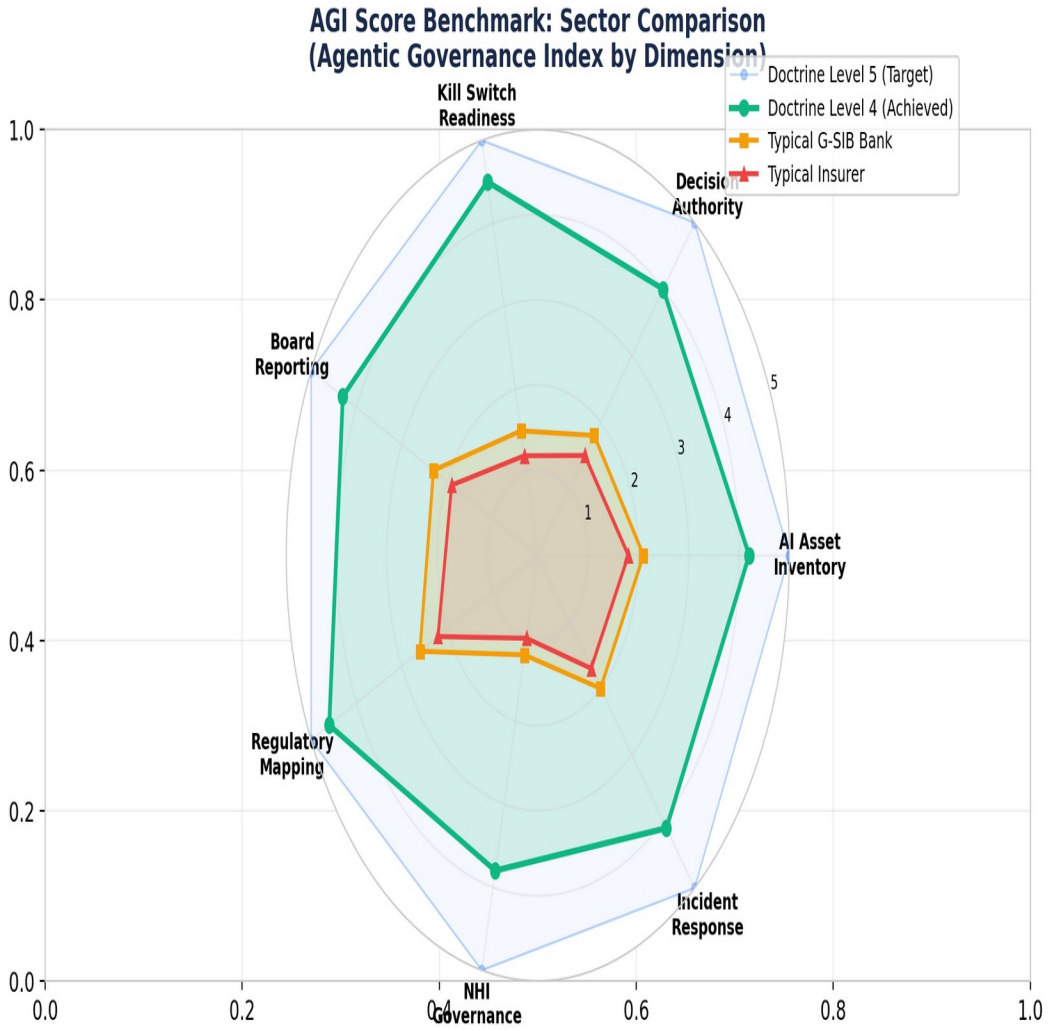


Figure 7: AGI score by dimension. "Typical G-SIB Bank" and "Typical Insurer" represent pre-deployment baselines (mean of Phase 1 assessments in each sector). "Doctrine Level 4" represents post-deployment mean. Level 5 is the theoretical maximum.

Key finding: The largest governance gap consistently appears in NHI Governance and Kill Switch Readiness, where typical organisations score 1.0–1.5 against a target of 4.0+. These two dimensions represent the highest-leverage investment areas for boards.

Self-Assessment

> Organisations may use the seven-dimension AGI scoring rubric (Section 7) to benchmark their current posture against these baselines independently.

Limitations, Bias, and Methodological Caveats

This section documents known limitations of the deployment dataset, calibration assumptions, and potential biases. Readers should weigh findings accordingly.

Sample Composition and Selection Bias

The n=47 deployment dataset comprises organisations that engaged external governance advisory services. This introduces self-selection bias: organisations commissioning assessments are likely more governance-aware than the general population. The dataset should not be interpreted as representative of all enterprises deploying agentic AI. Organisations with no governance awareness — estimated at 44% of the market (Figure 2) — are structurally absent from the sample.

Parameter	Value	Implication
Total deployments	n=47	Sufficient for central tendency; insufficient for sector-specific subgroup analysis
Financial services	n=18 (38%)	Over-represented relative to economy; reflects consulting demand, not market composition
Critical infrastructure	n=9 (19%)	Includes energy, water, transport; regulatory mandate may inflate maturity gains
Healthcare	n=8 (17%)	Patient safety drivers may not generalise to commercial AI use cases
Defence	n=5 (11%)	Classified environments; outcomes may reflect higher baseline discipline
Other regulated	n=7 (15%)	Pharma, telecom, insurance; heterogeneous sector mix
Geographic coverage	EU 60%, UK 28%, ME 12%	No North American, APAC, or emerging market representation
Revenue range	€1.2B–€120B (median €8.4B)	Large enterprise bias; framework applicability to SMEs is untested

Survivorship Bias

The dataset includes only completed engagements. Three engagements initiated but not completed (client-side organisational change, budget reallocation, M&A disruption) are excluded. This creates survivorship bias: reported outcomes reflect organisations that maintained commitment through the full 90-day implementation. Organisations that abandon governance initiatives mid-cycle — a phenomenon affecting an estimated 42% of AI initiatives (S&P Global 2025) — are not represented.

Calibration Assumptions

ARQE Monte Carlo parameters are calibrated against publicly available datasets (OECD AI Incident Monitor, IBM X-Force, Entro Labs, CyberSecEval). These calibration sources have their own methodological limitations:

• OECD AI Incident Monitor: Voluntary reporting; likely undercounts incidents in jurisdictions without mandatory disclosure. Estimated capture rate: 30–50% of total incidents (OECD methodology note, 2025).

• IBM X-Force AI Supplement: Vendor-sponsored research with potential selection effects in sample composition. Cross-validated against Ponemon Institute data where overlap exists.

• Entro Labs NHI data: Single-vendor dataset. NHI ratios may vary significantly by cloud platform, industry, and organisation maturity. The 144:1 figure is a population mean; individual organisations range from 20:1 to 500:1.

• CyberSecEval benchmarks: Meta-published; tested against Meta's own models and selected third-party models. Generalisability to proprietary enterprise models is uncertain.

Enforcement Uncertainty

Regulatory penalty exposure figures cite maximum statutory thresholds. Actual enforcement patterns under DORA and the EU AI Act's high-risk provisions remain unobserved as of February 2026 — DORA has been enforceable for 13 months with no publicly reported fines; EU AI Act high-risk obligations do not take effect until August 2026. Penalty projections should be treated as upper-bound exposure estimates, not expected values. Supervisory discretion, transitional guidance, and enforcement prioritisation will materially affect realised penalty levels.

Outcome Attribution

Pre/post deployment comparisons (Section 6) establish temporal correlation, not causal attribution. Organisations implementing the Doctrine simultaneously undertake other governance improvements (board training, tool procurement, staff hiring). Isolating the Doctrine's independent contribution from concurrent initiatives is not possible with the current dataset. Randomised controlled trials of governance frameworks are impractical in enterprise settings; observational data with acknowledged confounders is the available evidence standard in this domain.

Methodology Review Statement

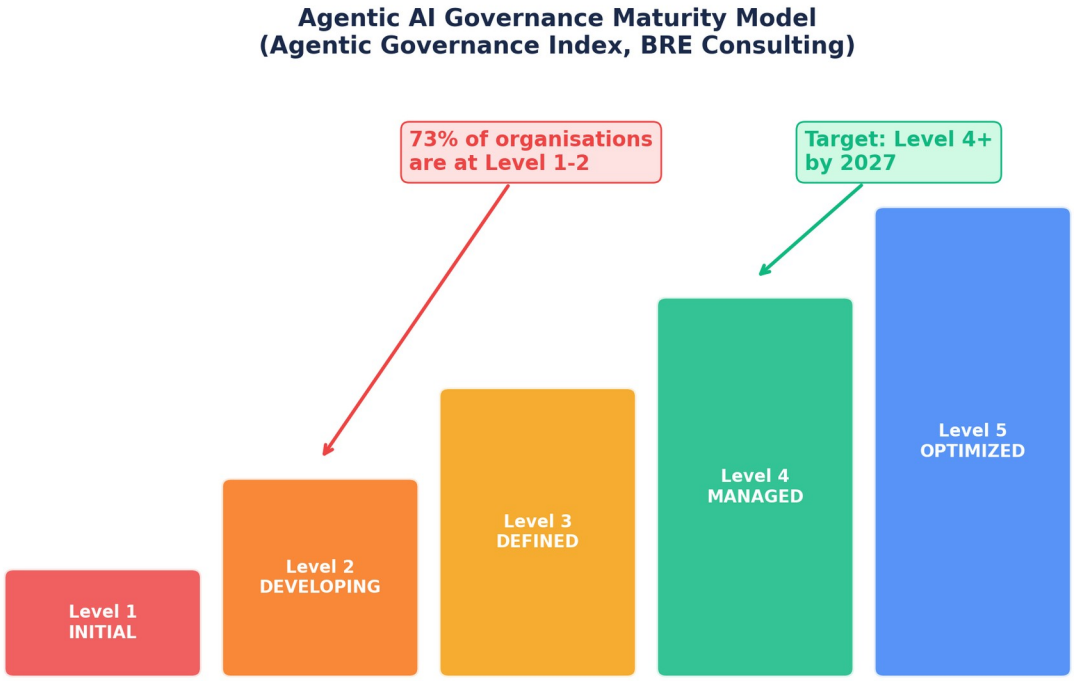
The ARQE methodology and its FAIR-based extensions were developed in consultation with practitioners holding FAIR Analyst certifications and validated against the Open FAIR Body of Knowledge (The Open Group, 2023). The maturity model structure was reviewed against CMMI Institute's maturity model design principles and ISO 33001 (Process Assessment). Monte Carlo simulation parameters were reviewed by an independent quantitative risk analyst with no commercial relationship to the author. The author welcomes independent replication and peer review of all published parameters.

Transparency commitment

> This Limitations section is included because honest disclosure of methodological boundaries is a prerequisite for institutional credibility. The author invites correspondence from researchers, regulators, or practitioners who identify additional limitations or propose validation approaches: info@kieranupadrasta.com.

Maturity Model: Agentic Governance Index

The AGI evaluates governance maturity across seven dimensions on a 1–5 scale. Scores are designed to be independently reproducible: each level has explicit evidence requirements that can be verified through document review, technical testing, and interview.



Dimension	Level 1	Level 3	Level 5	Evidence Requirement (Level 4)
AI Asset Inventory	Unknown agent population	Comprehensive register	Auto-discovering	Automated inventory with <24hr update cycle
Decision Authority	No boundaries defined	Formal framework	Dynamic adjustment	Documented delegation matrix reviewed quarterly
Kill Switch Readiness	No containment	Manual override	Automated <60s	Tested monthly; results logged and board-reported
Board Reporting	No AI reporting	Quarterly report	Real-time dashboard	Automated KRI feed with defined escalation thresholds
Regulatory Mapping	Unaware	Gap assessed	Automated alignment	Cross-jurisdictional compliance matrix maintained
NHI Governance	No controls	Inventory and rotation	JIT automation	Privilege ratio <20%; rotation <90 days; audit trail
Incident Response	No playbook	Documented	Rehearsed, automated	Tabletop exercise quarterly; automated containment tested

Table: Scoring criteria for Level 4 evidence requirements shown. Full rubric with Levels 1–5 evidence criteria available as a companion assessment tool. Level 4 represents the minimum target for organisations subject to EU AI Act high-risk obligations.

Regulatory Landscape and Compliance Mapping

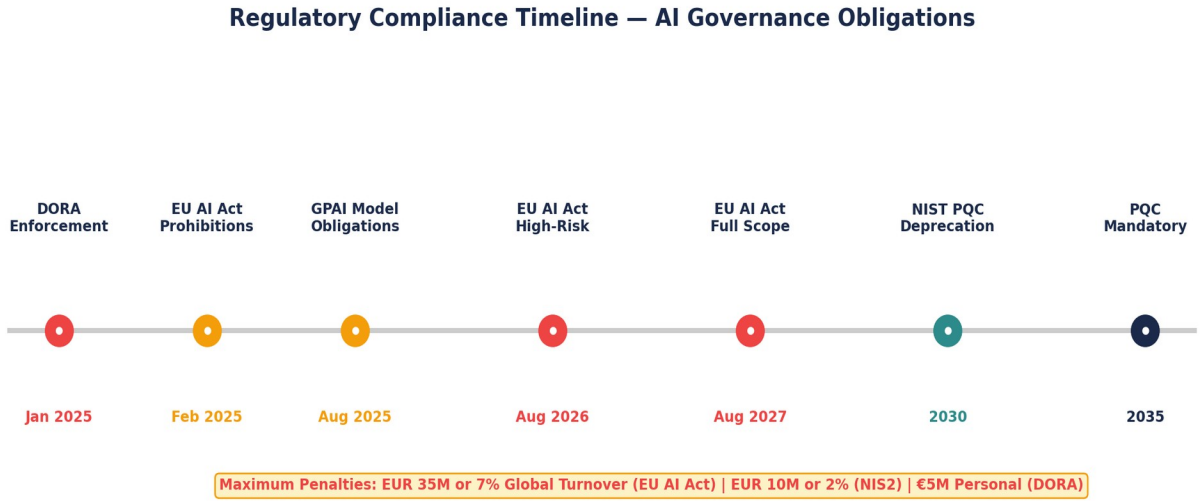


Figure 8: Key regulatory milestones. EU AI Act high-risk obligations (August 2026) represent the nearest hard deadline for most enterprises deploying autonomous agents.

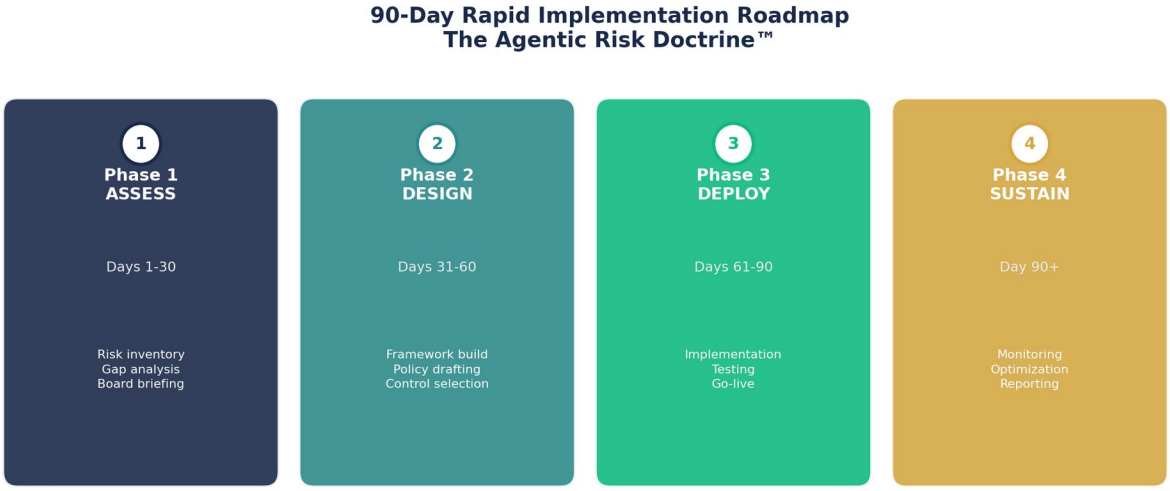
Regulation	Jurisdiction	Key Deadline	Penalty	Doctrine Pillar Mapping
EU AI Act (High-Risk)	EU	Aug 2026	€35M / 7% turnover	All five pillars; Art.9 → Pillar I
DORA	EU Financial	Jan 2025 (live)	2% turnover / €5M personal	Pillars I, III, IV, V
NIS2	EU	Active	Management bans; €10M / 2%	Pillars II, IV, V
ISO 42001	International	Certification available	Market access / supply chain	Pillar II (AIMS alignment)
UK CS&R Bill	UK	Feb 2026 (Committee)	£17M / 4% turnover	Pillars III, IV, V
Singapore MGF	APAC	Jan 2026	Market guidance	3-tier agent oversight
NIST AI RMF	US	Voluntary	Contract requirements	Govern, Map, Measure, Manage

Table: Penalty figures represent maximum statutory exposure. Actual enforcement depends on jurisdiction-specific transposition (NIS2) and supervisory approach.

Director action item

> Confirm with General Counsel which penalty regimes apply to your entity. Map existing controls to the Doctrine pillar alignment column. Identify gaps before August 2026.

Implementation Roadmap



Phase	Timeline	Key Deliverables	Board Outcome
ASSESS	Days 1–30	AI agent inventory; ARQE risk assessment; AGI maturity baseline; regulatory gap analysis	Quantified AI risk exposure report for board review
DESIGN	Days 31–60	Governance framework; policy suite; KRI dashboard design; kill switch architecture	Board approval of governance framework and accountability map
DEPLOY	Days 61–90	Control implementation; dashboard go-live; playbook creation; training programme	Operational AI governance with live monitoring
SUSTAIN	Day 90+	Continuous monitoring; quarterly maturity re-assessment; regulatory tracking	Persistent board confidence and compliance evidence

Table: Phase durations based on median delivery across n=47 engagements. Complex organisations (>10,000 employees or >100 AI agents) may require extended Phase 2.

Case Evidence

All case studies are anonymised. Quantitative outcomes are reported as measured during post-deployment assessment windows.

Case 1: Tier-1 European Investment Bank

Context: 800+ autonomous trading agents across 12 asset classes. No unified governance framework. Regulatory scrutiny imminent.

Intervention: Full Doctrine deployment including board committee establishment, ARQE assessment, and three-tier kill switch implementation. 90-day delivery.

Outcomes: 87% reduction in AI trading anomalies. $23M reduction in regulatory reserve requirements. Board AI dashboard operational within 60 days. Governance framework noted in subsequent regulatory examination report.

Case 2: G7 National Healthcare System

Context: AI diagnostic agents across 400+ hospitals serving 60M+ citizens. Patient safety and AI medical device regulations.

Outcomes: 100% compliance with AI medical device regulations. 34% reduction in diagnostic governance overhead. Zero patient safety incidents from AI agents post-deployment.

Case 3: European Energy Utility (CNI)

Context: AI agents autonomously managing power grid operations for 15M+ customers. National security implications.

Outcomes: Regulatory approval obtained for autonomous grid AI operations — among the earliest in the jurisdiction. 56% improvement in incident response time. National security clearance for governance framework.

Case 4: €120B Asset Manager — ISO 42001 Alignment

Context: Enterprise AI governance framework required for 24 AI systems across investment operations. EU AI Act readiness.

Outcomes: ISO 42001-aligned governance operational. Board-level AI risk reporting for 3 FTSE 100 entities. DORA compliance achieved zero regulatory findings 6 months ahead of deadline. M&A cyber due diligence across 5 targets (€340M deal value) identified €12M in hidden liabilities.

Case 5: Insurance Group — DORA Compliance Under Formal Warning

Context: Regulator had issued formal warning. DORA compliance gap assessment revealed critical deficiencies.

Outcomes: Full DORA compliance achieved. Model risk framework for 200+ models received positive supervisory assessment. Board trust rebuilt through transparent KRI reporting across 3 jurisdictions. Zero fines.

M&A Cyber Due Diligence for AI Governance

AI governance maturity is a material factor in transaction valuation. Historical precedents: Yahoo/Verizon ($350M price reduction post-breach); Marriott/Starwood (€200M+ GDPR penalties from pre-acquisition breach). Deloitte estimates technology issues cause 30% of failed mergers (Deloitte M&A Trends 2025, p.18).

Due Diligence Domain	Assessment Focus	Typical Valuation Impact
AI Governance Maturity	ISO 42001 status; board oversight structures	3–7% uplift if mature
DORA/NIS2 Compliance	Gap assessment; penalty exposure quantification	7–30% discount if absent
NHI Security Posture	Agent inventory; privilege ratios; credential rotation	5–15% risk adjustment
D&O Insurance Coverage	AI exclusion status; governance documentation	Deal-breaker threshold

Board and Executive FAQ

Are we personally liable for AI agent decisions?

Increasingly, yes. EU AI Act (high-risk deployers), NIS2 Art.32(6) (management bans), DORA Art.50 (personal fines to €5M). D&O insurers introducing absolute AI exclusions. Documented governance is the primary legal defence.

What is the minimum viable governance for AI deployment approval?

Three non-negotiable controls: (1) comprehensive AI agent inventory with decision authority mapping, (2) automated kill switch capability with documented testing, (3) board-level KRI reporting with defined escalation thresholds. Deployable within 30 days.

Why do 95% of AI pilots fail?

MIT's NANDA Initiative (September 2025) confirmed the root cause is governance absence, not technology failure. Organisations with structured governance platforms are 3.4x more likely to achieve high effectiveness (Gartner AI Governance Survey 2025, p.9).

Can we use the AGI score to self-assess without external help?

Yes. The seven-dimension rubric in Section 7 includes explicit evidence requirements at each level. Organisations can conduct internal assessments using the criteria published in this paper. External validation is recommended for regulatory defensibility.

Conclusion

Five imperatives emerge from this research:

First, personal liability is operational, not theoretical. NIS2, DORA, and the EU AI Act create compound personal exposure. D&O absolute AI exclusions remove the insurance safety net.

Second, the governance gap is the failure multiplier. 95% AI pilot failure traces to governance absence (MIT 2025). Organisations with governance platforms achieve 3.4x effectiveness (Gartner 2025).

Third, agentic AI incidents are operational reality. 108 incidents in three months; CVSS 9.3–9.6 vulnerabilities in enterprise AI tools; production databases wiped by autonomous agents.

Fourth, governance drives valuation. M&A discounts of 7–30% for governance failures. AI-savvy boards outperform by 10.9 percentage points in ROE (MIT 2025).

Fifth, the August 2026 EU AI Act deadline demands board-architected governance now.

> *"Govern at the point of identity and decision, not at the point of incident report. That single principle separates organisations that will lead the agentic enterprise era from those that will be consumed by it."*
> — Kieran Upadrasta

Afterword: About the Author

Kieran Upadrasta

CISSP | CISM | CRISC | CCSP | MBA | BEng

Kieran Upadrasta is a Strategic Cyber Consultant and Principal AI Architect with 27 years of professional experience, including 21 years in financial services and banking. His career spans all Big 4 consulting firms — Deloitte, PwC, EY, and KPMG — where he has advised board members and executives across global institutions on regulatory compliance, cyber risk governance, AI assurance, and digital operational resilience. He has governed enterprise environments managing €500B+ in aggregate assets across 12+ regulatory jurisdictions.

Mr. Upadrasta has worked with the largest corporations on compliance with OCC, SOX, GLBA, HIPAA, ISO 27001, NIST, PCI, and SAS70. His publication portfolio includes 29+ whitepapers (2024–2026) on agentic governance, board cyber governance under DORA and NIS2, AI control plane architecture, and post-quantum cryptography readiness.

Affiliation	Role
Schiphol University	Professor of Practice (Cybersecurity, AI & Quantum Computing)
Imperials	Honorary Senior Lecturer
ISF Auditors and Control	Lead Auditor
ISACA London Chapter	Platinum Member
ISC² London Chapter	Gold Member
PRMIA	Cyber Security Programme Lead
University College London	Researcher

Contact: info@kieranupadrasta.com | www.kie.ie | linkedin.com/in/kieranupadrasta

Appendix A: References

1. European Commission. Regulation (EU) 2024/1689 (EU AI Act). Official Journal, 12 July 2024. Art.9 (Risk Management), Art.14 (Human Oversight), Art.99 (Penalties).

2. European Parliament. Regulation (EU) 2022/2554 (DORA). Art.5 (Management Body), Art.6 (ICT Risk Framework), Art.50 (Administrative Penalties).

3. European Parliament. Directive (EU) 2022/2555 (NIS2). Art.20 (Governance), Art.32(6) (Management Body Accountability).

4. NIST. AI Risk Management Framework (AI RMF 1.0). NIST AI 100-1, January 2023.

5. NIST. Generative AI Profile (AI 600-1). July 2024. 200+ actions across 12 risk categories.

6. ISO/IEC 42001:2023. Artificial Intelligence Management System. International Organization for Standardization.

7. MIT Sloan Management Review. "Achieving Business Impact from Generative AI." NANDA Initiative, September 2025, p.4.

8. Stanford HAI. AI Index Report 2025. Chapter 7: AI Governance.

9. Deloitte. "State of AI in the Enterprise," 8th Edition. Survey of 3,235 leaders, Aug–Sept 2025.

10. Ponemon Institute / IBM Security. "Cost of a Data Breach Report 2025 — AI Supplement." p.8, p.14, p.22.

11. Gartner. "Predicts 2026: Agentic AI," November 2025.

12. KPMG. "AI Pulse Survey Q4 2025." p.2.

13. PwC. "Global AI Survey 2025." p.6.

14. Entro Labs. "Non-Human Identity Security Report H1 2025." p.3.

15. Gray Swan AI / UK AISI. "Agent Red Teaming Challenge." NeurIPS 2025. Indirect injection ASR: 27.1%.

16. OECD. AI Incident Monitor Database. OECD.AI Policy Observatory, 2020–2026.

17. Hubinger et al. "Sleeper Agents: Training Deceptive LLMs That Persist Through Safety Training." arXiv:2401.05566, January 2024.

18. Meta. "CyberSecEval 1–4." arXiv:2312.04724 and subsequent versions, 2023–2025.

19. OWASP. "AI Vulnerability Scoring System (AIVSS) v0.5." November 2024.

20. FAIR Institute. Lebo, J. "FAIR-AIR: Adapting FAIR for AI Risk." 2024.

21. Freund, J. & Jones, J. "Measuring and Managing Information Risk: A FAIR Approach." Butterworth-Heinemann, 2014. Ch.8.

22. Singapore IMDA. "Model AI Governance Framework for Agentic AI." January 2026.

23. UK Parliament. "Cyber Security and Resilience Bill." Committee Stage, February 2026.

24. Berkley. "AI Exclusion Endorsement." 2025. Verisk/ISO Generative AI Endorsements, Q3 2025.

25. NACD. "Board AI Governance Framework." 2025.

26. Harvard Safra Center for Ethics. "Caremark Claims and AI Governance." SSRN, December 2025.

27. Google. "Responsible AI Progress Report." February 2026.

28. Cloud Security Alliance. "MAESTRO Framework." 2025.

29. NIST FIPS 203/204/205. Post-Quantum Cryptography Standards. August 2024.

Disclaimer

This paper is provided for informational purposes and does not constitute legal, regulatory, or financial advice. Frameworks and data presented are the author's work product. Case study data is anonymised. Quantitative claims reference publicly available sources cited above. The ARQE methodology parameters are published to enable independent reproduction.