WHITEPAPER

Governing the Agentic Enterprise

From Shadow AI to Autonomous Security

A Strategic Framework for Board-Level AI Agent Governance,
Machine Identity Security, and Regulatory Compliance

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

27 Years' Cyber Security Experience | Big 4 Consulting
21 Years Financial Services | AI Cyber Security Programme Lead

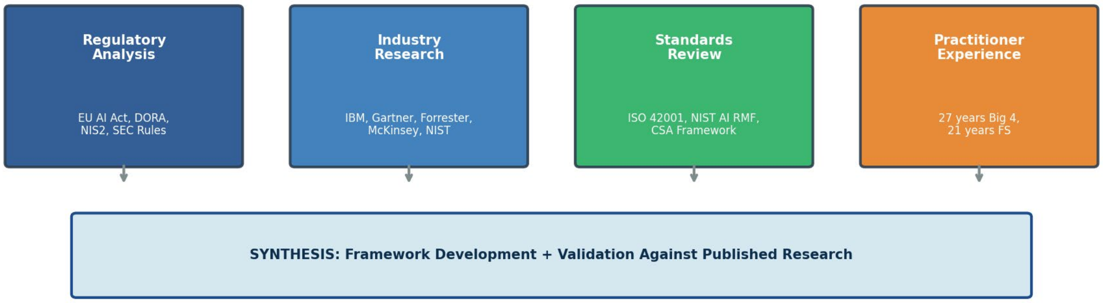
www.kie.ie | info@kieranupadrasta.com | January 2026

Table of Contents

Methodology and Evidence Standards

> This whitepaper synthesizes regulatory analysis, industry research, standards review, and 27 years of practitioner experience across Big 4 consulting and financial services. All statistics are cited with source, publisher, and date. Case studies are labeled as PUBLIC INCIDENT, COMPOSITE CASE (anonymized), or ILLUSTRATIVE SCENARIO.

Research Methodology


Regulatory Analysis
EU AI Act, DORA, NIS2, SEC Rules
Industry Research
IBM, Gartner, Forrester, McKinsey, NIST
Standards Review
ISO 42001, NIST AI RMF, CSA Framework
Practitioner Experience
27 years Big 4, 21 years FS
SYNTHESIS: Framework Development + Validation Against Published Research

Research conducted: Q4 2024 - Q1 2025 | Publication: January 2026

Evidence Classification

- **Primary Sources:** EU AI Act (2024/1689), DORA (2022/2554), NIS2 (2022/2555), NIST AI RMF
- **Industry Research:** IBM Cost of Data Breach 2025, Gartner Strategic Trends, Forrester Wave reports
- **Standards Bodies:** ISO/IEC 42001:2023, Cloud Security Alliance frameworks, IETF drafts
- **Validation:** All statistics traceable to published sources; directional claims noted

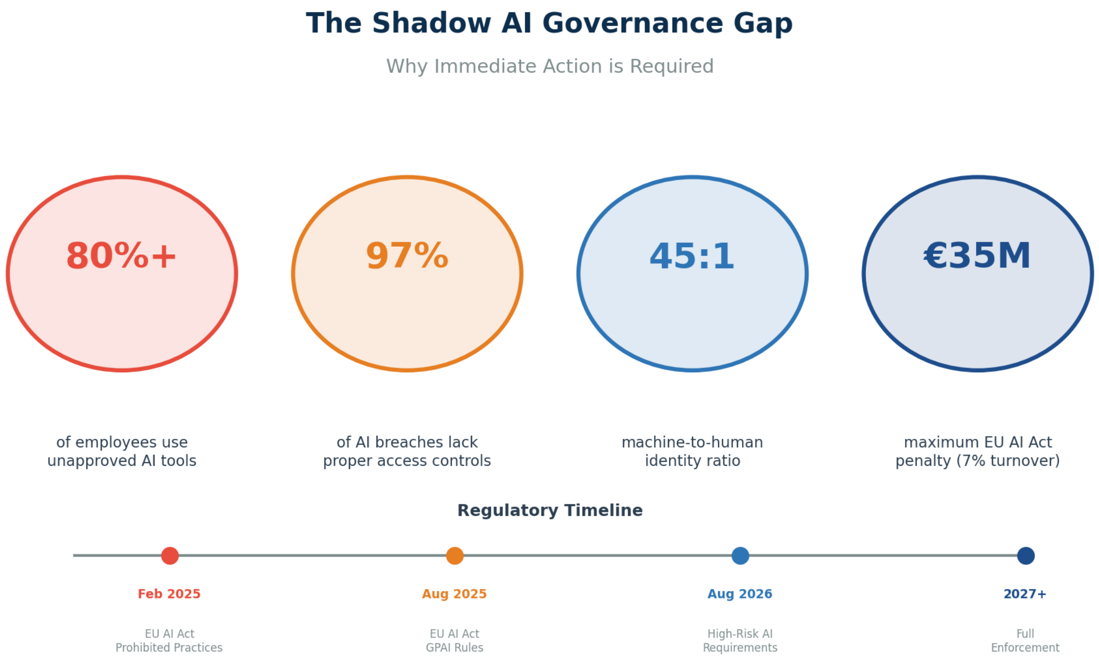
Executive Summary

> BOARD-LEVEL IMPERATIVE: Enterprises face material operational risk from ungoverned AI agents. Industry surveys indicate 80%+ of employees use unapproved AI tools,[1] shadow AI breaches add $670,000 in costs,[2] and EU AI Act penalties reach €35 million or 7% of global turnover.[3] This whitepaper provides an actionable governance framework.

The Governance Gap

The enterprise AI landscape has reached an inflection point. In large enterprises, machine identities often exceed human identities by an order of magnitude.[4] These agents execute transactions, provision infrastructure, and access sensitive data at machine speed—yet the majority operate outside governance frameworks.

The regulatory response has arrived. The EU AI Act (Regulation 2024/1689) imposes penalties up to €35 million or 7% of global annual turnover for prohibited AI practices—among the highest in EU regulatory history.[3] Board members face personal liability for governance failures under multiple jurisdictions.



Three Critical Findings

1. **Shadow AI presents material risk:** In industry surveys, over 80% of employees reported using unapproved AI tools.[1] IBM reports that 20% of organizations experienced shadow AI-related breaches in 2025, adding $670,000 in additional costs.[2]
2. **Machine identity governance requires modernization:** IBM found that among organizations experiencing AI breaches, the majority involved systems without proper

access controls.[5] Traditional IAM architectures were designed for humans—agents require fundamentally different approaches.

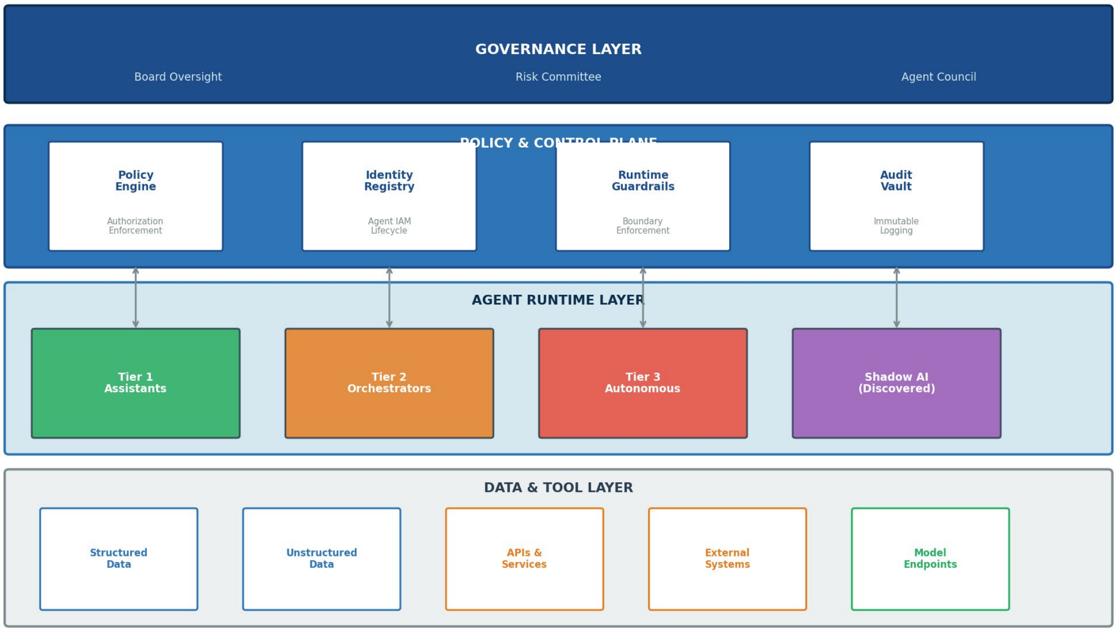
3. **Governance investment delivers measurable returns:** IBM reports organizations with AI security capabilities save $1.9 million per breach and detect incidents 80 days faster.[6] Gartner indicates organizations with comprehensive AI governance experience 40% fewer ethical incidents.[7]

1. Agent Governance Control Plane: Reference Architecture

> This reference architecture provides the foundational structure for enterprise AI agent governance. It integrates policy enforcement, identity management, runtime protection, and audit capabilities into a unified control plane.

Figure 1: Agent Governance Control Plane

Reference Architecture for Enterprise AI Agent Governance



Core Principles:

✓ Zero Trust: Every request verified
✓ Least Privilege: Task-specific access
✓ Continuous Monitoring: Real-time analytics
✓ Immutable Audit: Full provenance

1.1 Architecture Layers

4. **Governance Layer:** Board oversight, risk committee review, and Agent Council operational governance
5. **Policy & Control Plane:** Policy engine for authorization, identity registry for lifecycle management, runtime guardrails for boundary enforcement, audit vault for immutable logging
6. **Agent Runtime Layer:** Tiered agents (assistants, orchestrators, autonomous) plus discovered shadow AI under remediation
7. **Data & Tool Layer:** Structured/unstructured data, APIs, external systems, and model endpoints with classification controls

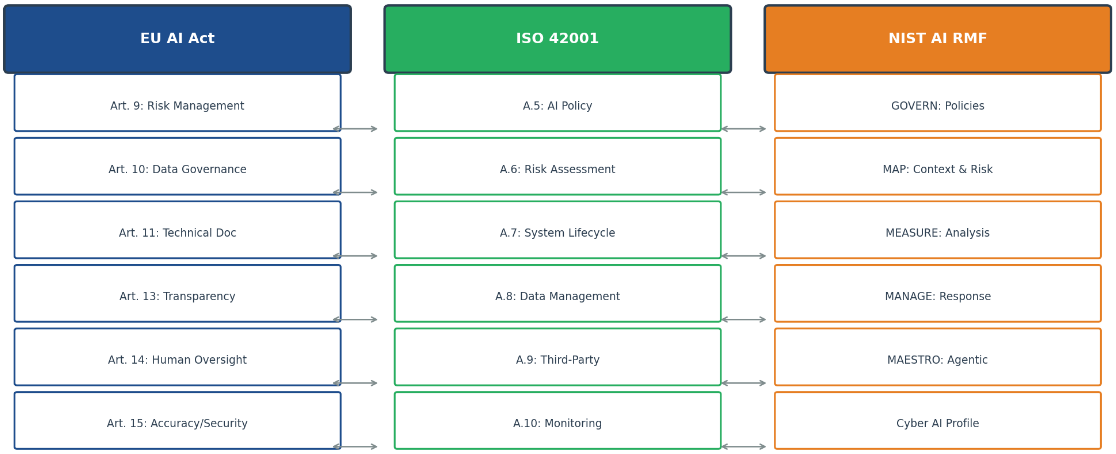
1.2 Control Plane Components

Component	Function	Key Capabilities
Policy Engine	Authorization enforcement	Risk-based decisions, boundary rules, escalation triggers
Identity Registry	Agent lifecycle management	Registration, classification, status tracking, kill switch
Runtime Guardrails	Boundary enforcement	Input validation, output filtering, tool restrictions
Audit Vault	Immutable logging	Action provenance, decision trails, compliance evidence

2. Regulatory Control Mapping: EU AI Act, ISO 42001, NIST

Figure 2: Regulatory Control Mapping Framework

EU AI Act ↔ ISO 42001 ↔ NIST AI RMF Alignment



Bidirectional arrows indicate control alignment opportunities

Organizations with ISO 27001 achieve ISO 42001 compliance 40% faster (industry survey)

2.1 EU AI Act (Regulation 2024/1689)

The EU AI Act establishes the first comprehensive legal framework for AI through risk-based classification. Penalties reach €35 million or 7% of global annual turnover for prohibited practices.[3]

Date	Requirement	Penalty Exposure
February 2025	Prohibited AI practices banned; AI literacy obligations[3]	€35M / 7% turnover
August 2025	General-purpose AI model obligations; penalty regime[3]	€15M / 3% turnover
August 2026	High-risk AI system requirements fully applicable[3]	Conformity assessment
August 2027	High-risk systems in regulated products compliant[3]	Full enforcement

2.2 ISO/IEC 42001:2023

ISO 42001 provides the first certifiable AI management system standard. It specifies 38 controls covering AI policy, risk evaluation, system lifecycle, and third-party oversight.[8] Industry experience suggests organizations with ISO 27001 certification achieve ISO 42001 compliance faster due to overlapping management system requirements.

2.3 NIST AI Risk Management Framework

NIST AI RMF provides four core functions: GOVERN (policies and accountability), MAP (context and capabilities), MEASURE (risk analysis), and MANAGE (response and monitoring).[9] The February 2025 MAESTRO Framework specifically addresses agentic systems through six analytical layers.[10]

2.4 Control Alignment Table

Governance Domain	EU AI Act	ISO 42001	NIST AI RMF
Risk Management	Art. 9	A.6	MAP, MEASURE
Data Governance	Art. 10	A.8	MAP
Technical Documentation	Art. 11	A.7	MAP, MANAGE
Transparency	Art. 13	A.5	GOVERN
Human Oversight	Art. 14	A.9	GOVERN, MANAGE
Accuracy & Security	Art. 15	A.10	MEASURE, MANAGE

3. Shadow AI: Quantifying the Governance Gap

> Shadow IT has challenged security teams for decades. Shadow AI represents an escalation: employees deploy AI agents with privileged access, creating concentration risk and control gaps that traditional security tools cannot detect.

3.1 Prevalence Evidence

Metric	Finding	Source
Employees using unapproved AI	80%+ in surveyed organizations	UpGuard, November 2025[1]
Organizations with shadow AI breaches	20%	IBM Cost of Data Breach 2025[2]
Additional breach cost	+$670,000	IBM Cost of Data Breach 2025[2]
Personal AI account usage	90%+ of companies have workers using personal chatbots	MIT Project NANDA[11]
Security professionals using shadow AI	Approximately 90%	UpGuard, November 2025[1]

3.2 Case Studies

> CASE LABELING: Cases below are classified as PUBLIC INCIDENT (documented, sourced), COMPOSITE CASE (anonymized from multiple engagements), or ILLUSTRATIVE SCENARIO (hypothetical based on threat modeling).

PUBLIC INCIDENT: Samsung ChatGPT Data Leak (2023)

Three separate incidents within 20 days at Samsung's semiconductor division exposed proprietary source code, chip testing sequences, and meeting transcription data when engineers uploaded confidential information to ChatGPT. Consequence: company-wide ban on generative AI tools.[12]

PUBLIC INCIDENT: Arup Deepfake Video Fraud (2024)

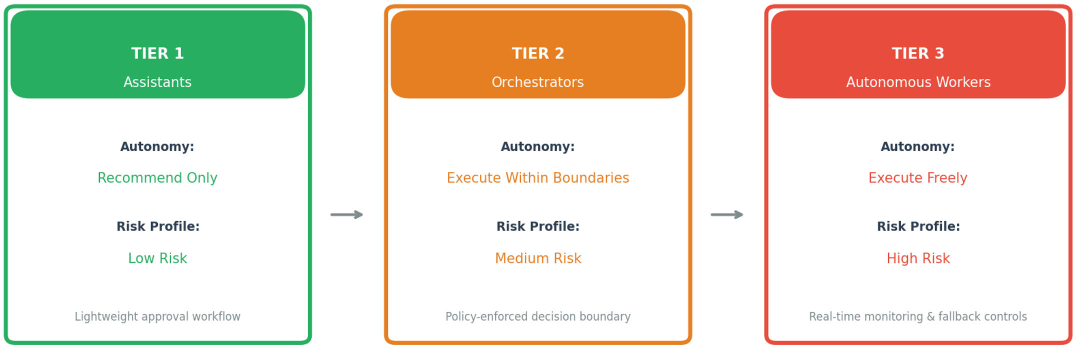
AI-generated deepfake video impersonating executives during a video conference call facilitated fund transfers totaling approximately HK$200 million (~US$25 million).[13] This demonstrates evolution from text-based social engineering to multimodal attacks.

ILLUSTRATIVE SCENARIO: Agent Prompt Injection

Scenario based on threat modeling: An AI agent trained to summarize customer support tickets is manipulated via prompt injection to extract PII and forward to an external API. Traditional DLP tools cannot parse natural language outputs, allowing exfiltration to continue for weeks before detection through anomalous network traffic analysis.

4. Agent Risk Classification and Scoring Framework

Agent Autonomy Levels and Governance Requirements



TIER 1
Assistants

Autonomy:
Recommend Only

Risk Profile:
Low Risk

Lightweight approval workflow

TIER 2
Orchestrators

Autonomy:
Execute Within Boundaries

Risk Profile:
Medium Risk

Policy-enforced decision boundary

TIER 3
Autonomous Workers

Autonomy:
Execute Freely

Risk Profile:
High Risk

Real-time monitoring & fallback controls

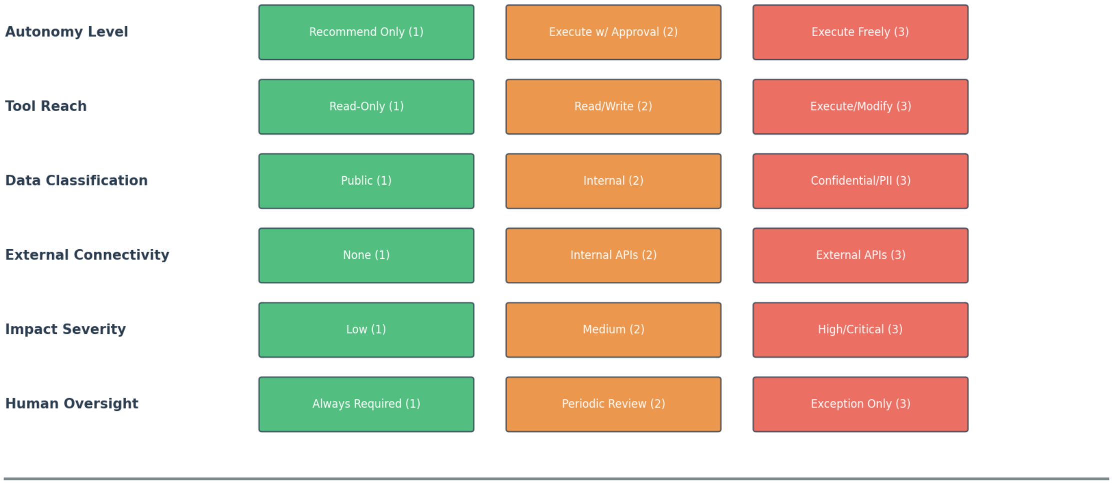
Governance proportionate to risk enables both speed and safety

4.1 Risk Scoring Methodology

Governance frameworks must distinguish between autonomy levels and apply proportionate oversight. The following scoring methodology enables consistent classification:

Figure 3: Agent Risk Scoring Framework

Risk Classification Dimensions and Control Requirements



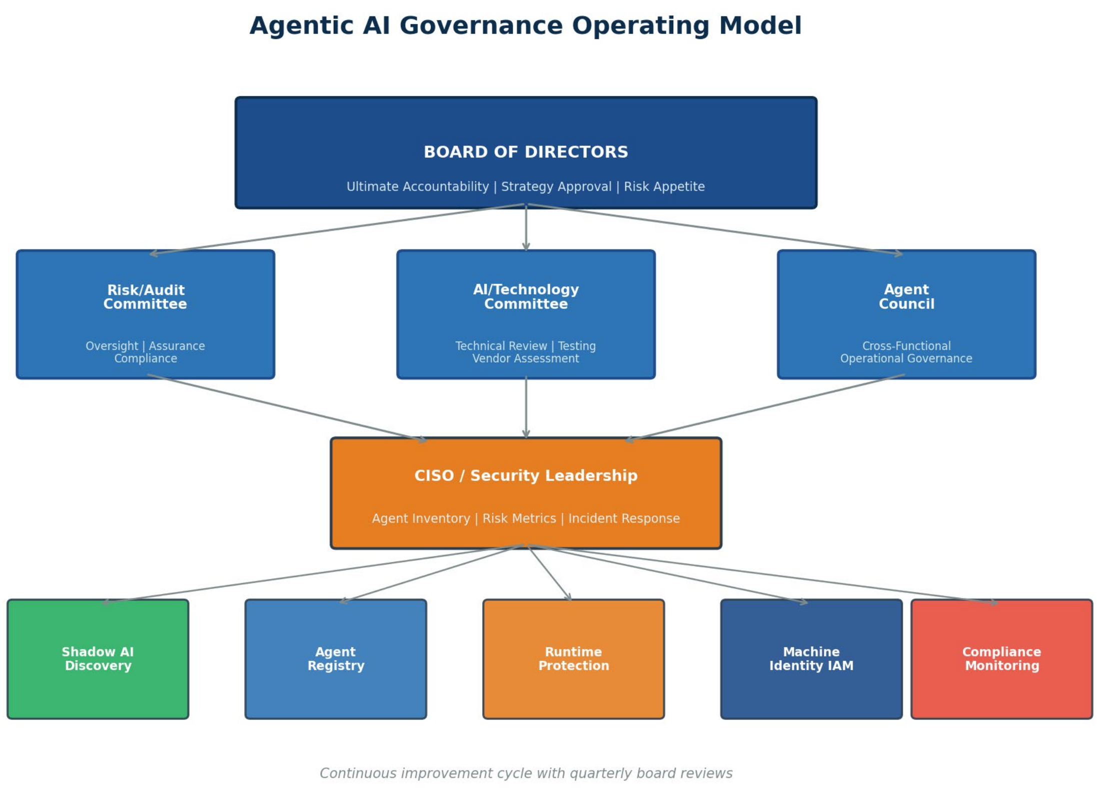
Autonomy Level	Recommend Only (1)	Execute w/ Approval (2)	Execute Freely (3)
Tool Reach	Read-Only (1)	Read/Write (2)	Execute/Modify (3)
Data Classification	Public (1)	Internal (2)	Confidential/PII (3)
External Connectivity	None (1)	Internal APIs (2)	External APIs (3)
Impact Severity	Low (1)	Medium (2)	High/Critical (3)
Human Oversight	Always Required (1)	Periodic Review (2)	Exception Only (3)

Total Score: 6-9 = Tier 1 (Low) | 10-14 = Tier 2 (Medium) | 15-18 = Tier 3 (High)

4.2 Control Requirements by Tier

Control	Tier 1 (Low)	Tier 2 (Medium)	Tier 3 (High)
Approval Process	Lightweight workflow	Business owner sign-off	Risk committee review
Monitoring	Periodic audit	Real-time alerts	Continuous + behavioral
Access Controls	Role-based	Attribute-based	Just-in-time + MFA
Kill Switch	Manual disable	4-hour SLA	Automatic triggers
Audit Retention	90 days	1 year	7 years

5. Agent Governance Operating Model


Agentic AI Governance Operating Model
BOARD OF DIRECTORS
Ultimate Accountability | Strategy Approval | Risk Appetite
Risk/Audit Committee
Oversight | Assurance Compliance
AI/Technology Committee
Technical Review | Testing Vendor Assessment
Agent Council
Cross-Functional Operational Governance
CISO / Security Leadership
Agent Inventory | Risk Metrics | Incident Response
Shadow AI Discovery
Agent Registry
Runtime Protection
Machine Identity IAM
Compliance Monitoring
Continuous improvement cycle with quarterly board reviews

5.1 Agent Council Charter

> The Agent Council provides operational governance through cross-functional oversight. The council should include: AI/ML Lead, Risk/Compliance Officer, Product Owner, Security Representative, and Executive Champion.

Charter Element	Specification
Mandate	Classify, approve, and monitor AI agents across the enterprise
Decision Rights	Approve Tier 2 agents; escalate Tier 3 to Risk Committee
Quorum	Security + Risk + Business representative minimum
Escalation Path	Unresolved issues to Risk Committee within 5 business days

Meeting Cadence	Weekly ops review / Monthly control assessment / Quarterly board pack

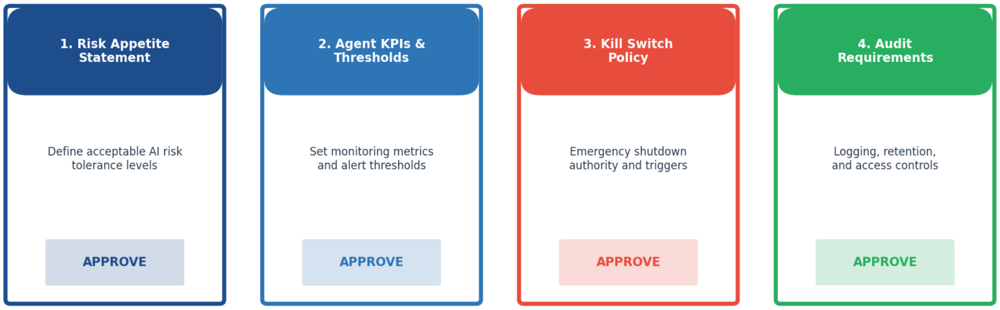
5.2 RACI Accountability Matrix

Activity	Board	Council	CISO	CRO
AI Strategy Approval	A	R	C	C
Risk Appetite Definition	A	C	R	R
Agent Classification	I	A	R	C
Runtime Policy Enforcement	I	C	A	C
Kill Switch Authorization	I	C	A	I
Board Training	R	C	R	R

6. Board Decision Pack: Required Approvals This Quarter

Figure 4: Board Decision Pack - This Quarter

Required Approvals for AI Agent Governance


1. Risk Appetite Statement
Define acceptable AI risk tolerance levels
APPROVE
2. Agent KPIs & Thresholds
Set monitoring metrics and alert thresholds
APPROVE
3. Kill Switch Policy
Emergency shutdown authority and triggers
APPROVE
4. Audit Requirements
Logging, retention, and access controls
APPROVE

Recommended: Complete all approvals within 90 days of program initiation

Board minutes should document: discussion summary, voting outcome, dissenting views

6.1 Decision 1: AI Risk Appetite Statement

The board must define acceptable AI risk tolerance levels, including: maximum acceptable shadow AI percentage, required compliance thresholds for high-risk systems, and quantified risk exposure limits. This statement anchors all subsequent governance decisions.

6.2 Decision 2: Agent KPIs and Thresholds

Approve specific metrics and alert thresholds: agent inventory completeness target, shadow AI detection rate, mean time to detect anomalies, compliance score requirements, and training completion minimums. These KPIs enable objective performance monitoring.

6.3 Decision 3: Kill Switch Policy

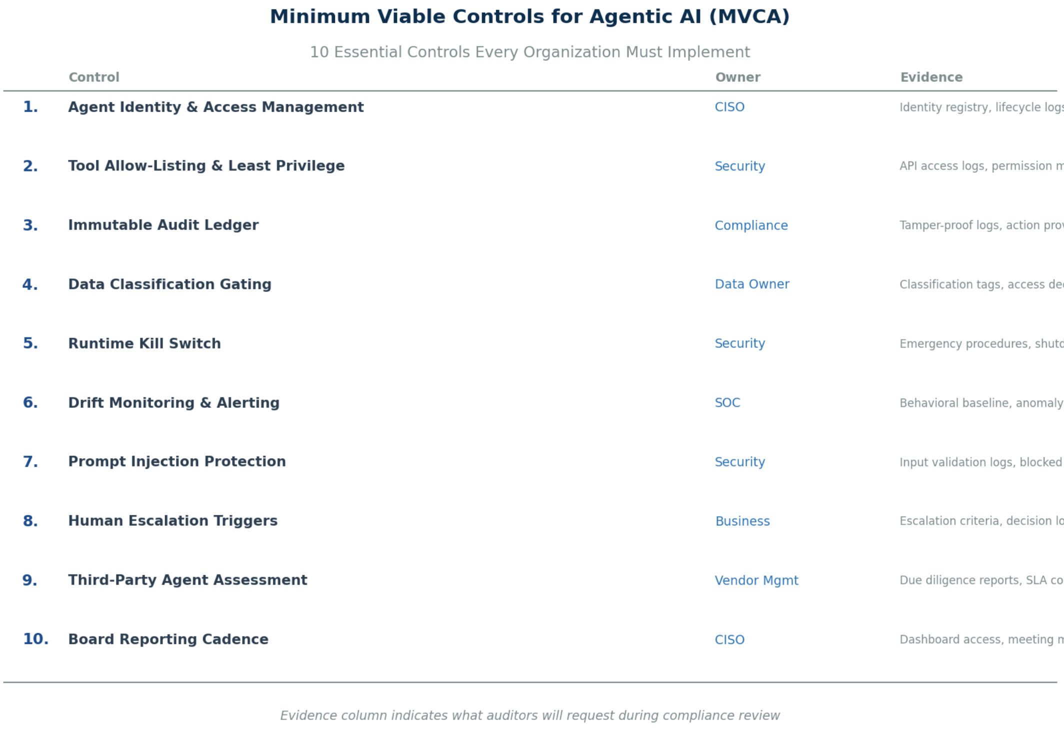
Approve emergency shutdown authority, including: who can invoke (CISO, Security Lead, Agent Council), under what triggers (breach detection, regulatory order, material risk event), and required response times by tier (Tier 3: immediate; Tier 2: 4 hours; Tier 1: 24 hours).

6.4 Decision 4: Audit Requirements

Define logging, retention, and access controls: all agent actions logged immutably, retention periods by tier (7 years for Tier 3), access restricted to authorized personnel, and compliance evidence preserved for regulatory examination.

7. Minimum Viable Controls for Agentic AI (MVCA)

> These 10 controls represent the baseline every organization must implement. Evidence column indicates what auditors will request during compliance review.



Minimum Viable Controls for Agentic AI (MVCA)

10 Essential Controls Every Organization Must Implement

	Control	Owner	Evidence
1.	Agent Identity & Access Management	CISO	Identity registry, lifecycle logs
2.	Tool Allow-Listing & Least Privilege	Security	API access logs, permission matrix
3.	Immutable Audit Ledger	Compliance	Tamper-proof logs, action provenance
4.	Data Classification Gating	Data Owner	Classification tags, access decisions
5.	Runtime Kill Switch	Security	Emergency procedures, shutdown logs
6.	Drift Monitoring & Alerting	SOC	Behavioral baseline, anomaly alerts
7.	Prompt Injection Protection	Security	Input validation logs, blocked attempts
8.	Human Escalation Triggers	Business	Escalation criteria, decision logs
9.	Third-Party Agent Assessment	Vendor Mgmt	Due diligence reports, SLA compliance
10.	Board Reporting Cadence	CISO	Dashboard access, meeting minutes

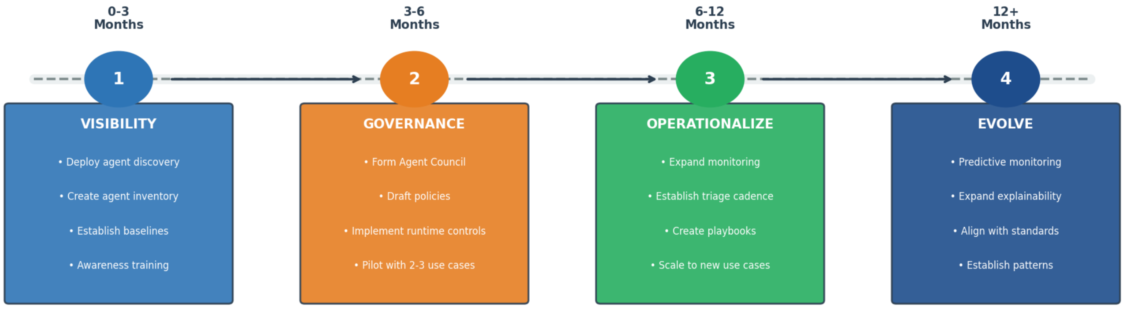
Evidence column indicates what auditors will request during compliance review

#	Control	Owner	Evidence
1	Agent Identity & Access Management	CISO	Identity registry, lifecycle logs
2	Tool Allow-Listing & Least Privilege	Security	API access logs, permission matrix
3	Immutable Audit Ledger	Compliance	Tamper-proof logs, action provenance
4	Data Classification Gating	Data Owner	Classification tags, access decisions
5	Runtime Kill Switch	Security	Emergency procedures, shutdown logs
6	Drift Monitoring & Alerting	SOC	Behavioral baseline, anomaly alerts

7	Prompt Injection Protection	Security	Input validation logs, blocked attempts
8	Human Escalation Triggers	Business	Escalation criteria, decision logs
9	Third-Party Agent Assessment	Vendor Mgmt	Due diligence reports, SLA compliance
10	Board Reporting Cadence	CISO	Dashboard access, meeting minutes

8. Implementation Roadmap with Exit Criteria

Implementation Roadmap to Governed Agentic Enterprise



Success: Zero material incidents • Regulatory readiness • Scaled adoption across business units

8.1 Phase 1: Visibility (Months 1-3)

Deliverables	Exit Criteria	Success Metrics
Agent registry	95% of agents registered	Registry completeness rate
Tool inventory	100% prod agents in audit logging	Logging coverage rate
Risk tiering	All agents classified by tier	Classification completion
Awareness training	80% staff completion	Training completion rate

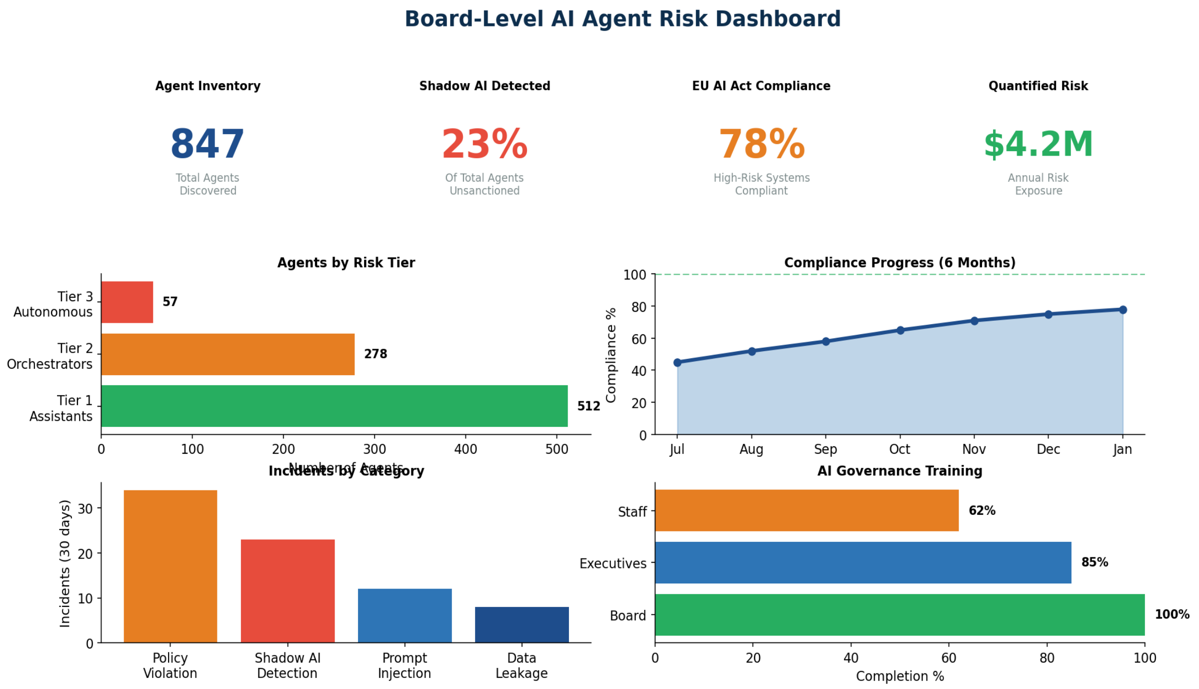
8.2 Phase 2: Governance (Months 3-6)

Deliverables	Exit Criteria	Success Metrics
Agent Council	Charter approved, first meeting held	Council meeting cadence
Policies	Data access, approval, security policies signed	Policy coverage
Runtime controls	IAM, DLP, monitoring deployed	Control deployment rate
Pilot validation	2-3 use cases through full governance cycle	Governance latency

8.3 Phase 3: Operationalize (Months 6-12)

Deliverables	Exit Criteria	Success Metrics
Full monitoring	All agents under active monitoring	Monitoring coverage
Triage cadence	Weekly review operational	Exceptions cleared rate
Playbooks	Documented for each issue type	Incident response time
Expansion	All business units onboarded	Organizational coverage

9. Board-Level AI Agent Dashboard



9.1 Governance KPIs

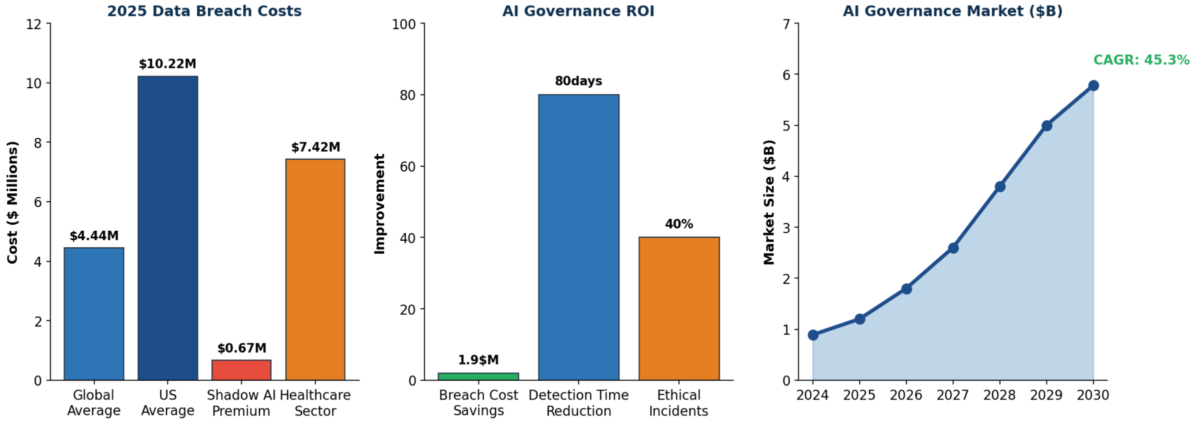
KPI	Description	Target
Agent Inventory	Total agents discovered and classified	100% visibility
Shadow AI Rate	Percentage of unsanctioned agents	<10%
Compliance Score	EU AI Act / ISO 42001 adherence	>90%
Tier Distribution	Agents by risk classification	Per risk appetite
Policy Violations	Monthly governance breaches	<5 critical
Training Completion	Board/executive AI training	100% board

9.2 Operational Metrics

KPI	Description	Target
MTTD	Mean Time to Detect agent anomalies	<24 hours

MTTR	Mean Time to Respond/Contain	<4 hours
Approval Latency	Time from request to deployment	<3 business days
False Positive Rate	Alerts requiring no action	<5%
Audit Coverage	Agents with complete trails	100%
Risk Quantification	Annual exposure in currency	Board-defined

10. Economic Value Model



2025 Data Breach Costs

AI Governance ROI

AI Governance Market ($B)

10.1 Cost of Inaction

Risk Category	Potential Impact	Source
Shadow AI breach premium	+$670,000 per incident	IBM 2025[2]
EU AI Act maximum penalty	€35M or 7% global turnover	EU Regulation[3]
Average US breach cost	$10.22 million	IBM 2025[2]
Healthcare sector average	$7.42 million	IBM 2025[2]

10.2 Return on Governance Investment

Benefit	Impact	Source
AI security deployment savings	$1.9M per breach avoided	IBM 2025[6]
Detection time reduction	80 days faster	IBM 2025[6]
Ethical incident reduction	40% fewer with governance	Gartner[7]

10.3 Value Model Template

> Conservative ROI Calculation: Cost of one shadow AI incident: $4.44M + $0.67M premium = $5.11M Expected reduction with governance: 40% (conservative) Potential savings per incident avoided: $2.04M Governance implementation cost (estimate): $500K-$1.5M Payback period: Single incident avoided

11. What Governance Cannot Prevent

> Governance reduces probability and severity of incidents—it does not eliminate risk. Mature organizations acknowledge limitations and design for resilience rather than prevention alone.

Even with comprehensive governance, organizations remain exposed to:

- **Insider misuse:** Authorized users may circumvent controls or misuse access privileges
- **Misconfiguration:** Human error in policy definition or control deployment
- **Third-party vendor failures:** Dependencies on external AI platforms and services
- **Model hallucination:** AI outputs may be inaccurate despite proper governance
- **Zero-day vulnerabilities:** Novel attack vectors not yet addressed by controls
- **Regulatory evolution:** Standards continue to develop; compliance today may require adjustment

Mitigation: Design for resilience, maintain incident response capabilities, conduct regular tabletop exercises, and maintain cyber insurance appropriate to AI risk exposure.

Appendix A: Board Governance Checklist

Governance Item	Status
Board-approved AI risk appetite documented	□
Agent inventory complete with classification	□
Agent Council established and operational	□
Shadow AI discovery tools deployed	□
Kill switch policy approved	□
Runtime guardrails implemented	□
Audit logging active for all agents	□
Board training completed (EU AI Act / ISO 42001)	□
Third-party AI vendor risk assessed	□
Incident response playbooks documented	□
Quarterly board reporting established	□
Compliance evidence repository maintained	□

About the Author

Kieran Upadrasta

CISSP, CISM, CRISC, CCSP | MBA | BEng

Kieran Upadrasta is a cyber security practitioner with 27 years of professional experience, including 21 years in financial services and banking. His career spans all four major consulting firms—Deloitte, PwC, EY, and KPMG—advising board members and senior executives on regulatory compliance, AI governance, and digital operational resilience.

Professional Memberships

- Honorary Senior Lecturer
- Lead Auditor, ISF Auditors and Control
- Platinum Member, ISACA London Chapter
- Gold Member, ISC² London Chapter
- Cyber Security Programme Lead, PRMIA
- Researcher, University College London (UCL)

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

Endnotes and Sources

[1] UpGuard, "The State of Shadow AI," November 2025. Survey of enterprise security posture indicating 80%+ employees use unapproved AI tools.

[2] IBM Security, "Cost of a Data Breach Report 2025," July 2025. Global study of 604 organizations. Shadow AI breach finding: 20% of organizations, +$670K cost premium.

[3] EU AI Act, Regulation (EU) 2024/1689, Official Journal of the European Union, July 12, 2024. Articles 71, 99-101 (penalties).

[4] CyberArk, "2024 Identity Security Threat Landscape," 2024. Machine-to-human identity ratios in enterprise environments.

[5] IBM Security, "Cost of a Data Breach Report 2025." Finding on access control gaps in AI breaches.

[6] IBM Security, "Cost of a Data Breach Report 2025." Organizations with AI security saved $1.9M, detected 80 days faster.

[7] Gartner, "Top Strategic Technology Trends 2025," October 2024. Prediction on AI governance and ethical incident reduction.

[8] ISO/IEC 42001:2023, "Information technology — Artificial intelligence — Management system," December 2023.

[9] NIST, "AI Risk Management Framework 1.0," January 2023.

[10] NIST, "MAESTRO Framework for Agentic AI," February 2025.

[11] MIT Sloan, "Project NANDA: State of AI in Business," 2025. Survey of generative AI adoption patterns.

[12] TechCrunch, "Samsung bans ChatGPT after code leak," May 2, 2023.

[13] CNN, "Finance worker pays out $25 million after video call with deepfake CFO," February 4, 2024.

This whitepaper is for informational purposes only and does not constitute legal advice. Organizations should consult qualified legal counsel for specific compliance guidance.