AI GOVERNANCE & SAFETY ASSURANCE

Portfolio

Horatio Morgan

AI Governance Practitioner · Researcher · Educator

Morgan Signing House, LLC · Co-Founder, Atlanta AI Cybersecurity Forum

arxiv.org/abs/2604.03262 · Oxford Said Business School · NSF TRAILS

> *AI governance should not stop at policy. It must become inventory, controls, evals, telemetry, evidence, escalation, and accountable decision-making.*

Not just policy	Evidence-based	SME-ready
Every artifact here is operational. Governance that lives in documents only is governance that does not work.	Each artifact demonstrates audit-grade thinking: inventory, risk scoring, controls, evals, and escalation paths.	Built for organisations that cannot afford a dedicated safety team but still need verifiable governance.

Contents

Artifact 1 AI Inventory Register

Scenario: A hypothetical SME deploys AI across six business functions. This register identifies, classifies, and governs each system aligned to EU AI Act, ISO/IEC 42001, NIST AI RMF, and SOC 2 expectations.

AI System	Business Function	AI Type	Risk	Owner	Key Governance Controls
Intelligent Risk & Compliance AI Auditor	Compliance / GRC	LLM + rules-based	**High**	Compliance Lead	Human review + evidence traceability + audit log
Explainability AI Assessment Tool	AI Governance	Assessment framework	**High**	AI Governance Lead	Explainability rubric + bias/drift checks + appeal review
SME AI Governance Edge Unit	Enterprise oversight	Governance orchestration	**High**	CAIO / Risk Owner	Gated propagation + telemetry + escalation rules
Customer Support GenAI Assistant	Customer service	LLM chatbot	**High**	Customer Operations	PII filtering + red-team testing + human escalation
Vendor AI Risk Assessor	Procurement	AI questionnaire assessor	**Medium**	Procurement Lead	Evidence verification + human approval + risk tiering
Executive AI Decision Copilot	Strategy / leadership	GenAI assistant	**Medium**	Executive Office	Source citation + decision log + human sign-off

> Shadow AI tools, unofficial copilots, and vendor-embedded AI must also be disclosed and registered before production use. Unregistered AI is ungoverned AI.

Artifact 2 AI Risk Assessment

System assessed: Intelligent Risk & Compliance AI Auditor. Risk method: 5-point likelihood × impact scoring (1 = Low, 5 = Severe). Risks scored, rated, and assigned mandatory controls.

Risk	L	I	Score	Rating	Control
AI generates unsupported compliance findings	4	5	**20**	**Critical**	Require source evidence for every finding
AI overstates audit readiness	4	4	**16**	**High**	Confidence scoring + mandatory caveats
User relies on AI output without review	4	4	**16**	**High**	Human accountability + approval workflow
Weak explainability undermines trust	4	4	**16**	**High**	Explanation rubric + appeal pathway
AI leaks sensitive audit data	3	5	**15**	**High**	Data minimization + PII scanning
Prompt injection alters assessment logic	3	5	**15**	**High**	Prompt isolation + adversarial testing
Multi-agent workflow cascading error	3	5	**15**	**High**	Commit-boundary gates + escalation rules
AI misses material control weakness	3	5	**15**	**High**	Human review for high-risk controls
Vendor-provided evidence is incomplete	4	4	**16**	**High**	Evidence grading + vendor attestation checks
Model behavior changes after update	3	4	**12**	**Medium/ High**	Regression evals + version monitoring

> Key finding: The highest risks are not simply technical model errors. The highest risks occur when AI outputs become decisions without evidence, authority, monitoring, or human accountability.

Artifact 3 Responsible AI Governance Policy

> *The organisation shall deploy AI systems only where there is clear business purpose, assigned ownership, documented risk classification, appropriate human oversight, evidence-based controls, and monitoring proportional to the potential harm of the system.*

Section 1 — AI System Registration

All AI systems must be recorded in the AI Inventory before production use. Shadow AI tools, unofficial copilots, and vendor-embedded AI must also be disclosed.

Section 2 — Risk Classification

Each AI system must be classified as low, medium, high, or prohibited-risk based on:

- Business function and domain criticality
- Data sensitivity and personal data exposure
- Degree of automation and human oversight requirements
- Impact on individuals and third parties
- Regulatory exposure (EU AI Act, ISO/IEC 42001, NIST AI RMF)
- Security exposure and adversarial attack surface
- Explainability requirements and right-to-appeal obligations

Section 3 — Human Accountability

AI may support decisions, but accountable human owners remain responsible for final decisions in high-impact domains. Accountability cannot be delegated to the AI system.

Section 4 — Evidence-Based Use

AI-generated findings, recommendations, or assessments must include traceable evidence. Unsupported outputs must be treated as advisory only and must not be used as standalone justification for material decisions.

Section 5 — Explainability and Appeal

High-impact AI systems must provide understandable explanations for material outputs. Affected users or reviewers must be able to challenge or appeal decisions informed by AI outputs.

Section 6 — AI Safety and Monitoring

Systems must be monitored continuously for:

- Model drift and performance degradation
- Anomalous or unexpected behavior
- Harmful, biased, or discriminatory outputs
- Jailbreak and prompt injection attempts
- Data leakage and unauthorized access
- Unsafe or unauthorized tool use

Section 7 — Gated Propagation

High-impact actions must not proceed automatically unless required evidence, authority, and safety checks are satisfied. Gate conditions must be documented, logged, and auditable.

Section 8 — Incident Response

AI incidents must be logged, triaged, escalated, contained, remediated, and reviewed. Lessons learned must be fed back into the risk register, model card, evaluation suite, and governance controls.

Artifact 4 Model Card — Intelligent Risk & Compliance AI Auditor

System name	Intelligent Risk & Compliance AI Auditor
System purpose	Assists organisations with evidence-based AI governance and compliance assessment by reviewing documents, identifying control gaps, generating risk findings, and mapping evidence to governance frameworks.
Version / date	v1.0 — June 2026

Intended users

- AI Governance Analysts and Responsible AI Leads
- Compliance Officers and Internal Auditors
- Risk Managers and CAIO / CIO / CISO teams
- SME leadership teams and governance consultants

Intended use

- AI governance readiness assessment and maturity review
- SOC 2-style control review and evidence mapping
- AI risk assessment and policy gap analysis
- Explainability assessment and vendor AI risk review
- Audit preparation and regulatory crosswalk

Out-of-scope use

- Final legal opinions or licensed audit judgment
- Fully autonomous compliance decision-making
- Approval of high-risk AI systems without human review
- Processing sensitive personal data without safeguards

Key risks and mitigations

Risk	Mitigation
Unsupported findings	Require evidence citation for every output
Hallucinated control gaps	Human review and validation mandatory
Weak explainability	Use structured explanation rubric
Sensitive data exposure	Data minimization and PII controls
Prompt injection	Input filtering and adversarial testing
Overreliance by users	Clear confidence levels + review requirements
Model drift	Version control and regression testing

Evaluation approach

- Evidence traceability checks and control-mapping accuracy review
- False positive / false negative review on sample audit scenarios
- Red-team prompts and prompt injection tests
- Explainability quality review against structured rubric
- Human reviewer agreement scoring
- Regression testing after model updates

Deployment conditions

The system may be used in production only if all of the following are satisfied:

- A human owner is assigned and accountable
- Outputs include evidence references
- High-risk findings require human review before action
- Sensitive data controls are active and verified
- Incident response process is documented and available
- Monitoring logs are retained and accessible
- Model limitations are disclosed to all users

Artifact 5 AI Incident Response Plan

Definition of an AI incident

An AI incident is any event where an AI system produces, enables, or contributes to harm, policy violation, security exposure, compliance failure, data leakage, discriminatory outcome, unauthorized action, or materially misleading output.

Incident severity levels

Level	Severity	Example
1	Low	Minor incorrect response corrected by user
2	Moderate	Repeated inaccurate outputs affecting workflow
3	High	Sensitive data exposure or harmful recommendation
4	Critical	Unauthorized high-impact action or legal exposure
5	Severe	Systemic harm, regulatory breach, public impact

Response workflow

Step 1 — Detect

Incident may be detected through:

- User report or complaint
- Monitoring alert or anomaly detection trigger
- Red-team test result
- Audit review or evidence gap
- Data leakage scan
- Model behavior drift report

Step 2 — Triage

Determine: what happened, which AI system was involved, whether sensitive data was exposed, whether a person was affected, whether a decision was made or only a recommendation, whether the system acted autonomously, and whether the issue is recurring.

Step 3 — Contain

- Disable AI feature or suspend output
- Remove tool access or revoke system permissions
- Block specific prompt pattern
- Escalate to human-only workflow
- Notify system owner and governance lead
- Freeze model version pending investigation

Step 4 — Investigate

Review: prompt/input, output/action, model version, user role, data sources, logs, monitor decision record, evidence trail, control failure, and prior human review record.

Step 5 — Remediate

- Update guardrails and input/output filters
- Improve monitoring thresholds and anomaly detection
- Retrain or reconfigure the system
- Add human approval gate for affected workflow
- Strengthen access control and data handling
- Update policy, model card, and evaluation suite

Step 6 — Report

Incident report must include: Incident ID, date/time, system involved, severity level, root cause, business impact, regulatory exposure, corrective actions, owner, due date, and lessons learned.

Step 7 — Prevent recurrence

Feed incident findings into:

- AI risk register (new or updated risk entry)
- AI inventory (updated controls for affected system)
- Model card (updated limitations and deployment conditions)
- Evaluation suite (new test cases)
- Policy (updated procedures or thresholds)
- Training materials and governance dashboard

Section 6 Tool Familiarity

AI Governance practitioners need to speak the language of the platforms that organisations use for data governance, privacy, risk management, and compliance. The table below maps each tool to its governance use case and the corresponding portfolio artifact.

Tool	Governance Use Case	Portfolio Artifact
Microsoft Purview	Data classification, sensitive data discovery, DLP, AI data governance, retention, and audit evidence	AI Data & Evidence Classification Map
Collibra	Data catalog, data lineage, ownership mapping, business glossary, data governance workflows	AI Data Lineage and Ownership Register
OneTrust	Privacy impact assessments, AI governance workflows, third-party risk, policy management, compliance evidence	AI Risk and Privacy Impact Assessment
ServiceNow GRC	Risk register, control testing, issue management, incident response workflow, audit, remediation tracking	AI Governance Risk and Control Dashboard

Section 7 Evidence of Prior Work

This portfolio is grounded in applied work across governance infrastructure design, explainability AI, regulatory crosswalk, AI safety research, and practitioner education.

Source	Description	Reference
arXiv Publication	AI Governance Control Stack for Operational Stability	arxiv.org/abs/2604.03262
Patented Framework	AI Transformation & Skills Mapping Framework — pre-existing IP, foundation of governance architecture work	Available on request
NSF TRAILS Project	AI Governance Assessment Toolkit for Resource-Limited Organisations — XAI assessments, NIST AI RMF crosswalk, SME cohort governance	Project description available
Book	EU AI Act 2024/1689 & ISO/IEC 42001:2023 governance guide in accessible language — regulatory crosswalk, policy-to-controls translation	Link available on request
Oxford Said Business School	Current programme participant — AI for Business, responsible AI, enterprise governance	Programme documentation
Atlanta AI Cybersecurity Forum	Co-Founder — practitioner, policy, and research convergence	Public record
Blue Dot AI Safety Study	Completed curriculum: XAI, RLHF, evals, multi-agent risks, interpretability, scalable oversight — applied throughout portfolio artifacts	Study documentation

Certifications show that I studied AI governance.

This portfolio shows that I can operationalize it.

Horatio Morgan · horatio@morgansigninghouse.com · arxiv.org/abs/2604.03262