Agile Artificial Intelligence Governance: A Practical Approach to Responsible Corporate Adoption

Gustavo Adolfo Venegas Olivera
AI Security & Governance Research
linkedin.com/in/gvenegascc/

January 2026

Abstract

This paper presents an artificial intelligence (AI) governance framework illustrated through an anonymized large-enterprise scenario, designed to balance the critical need for accelerated innovation with the management of emerging risks, privacy, and cybersecurity. During the 2025–2026 transition, the scenario reflects the challenge of mass adoption of generative AI and foundation models. The proposed solution is an "Agile Governance" model that integrates international standards such as ISO/IEC 42001 and the ISMS Forum/Cisco guidelines, and that establishes an ecosystem of defined roles (CAIO, CISO, CDO) and an automated risk-classification process. The paper details guiding principles, a multidimensional validation lifecycle (technical, ethical, and functional), and an enabling reference tech stack (AI initiatives hub), providing a practical roadmap for organizations seeking to scale AI adoption safely and responsibly.

1 Introduction

The year 2025 marked a turning point in the global technology industry, with the Cambrian explosion of generative artificial intelligence (GenAI) and the proliferation of intelligent assistants (copilots). For large corporations, this advancement created an operational dichotomy: on the one hand, an unavoidable opportunity for cost optimization and customer-experience personalization; on the other, an emerging risk vector of unknown magnitude, characterized by potential intellectual-property leakage, the emergence of undetected algorithmic bias, and the *Shadow AI* phenomenon, in which employees adopt unsanctioned tools [3].

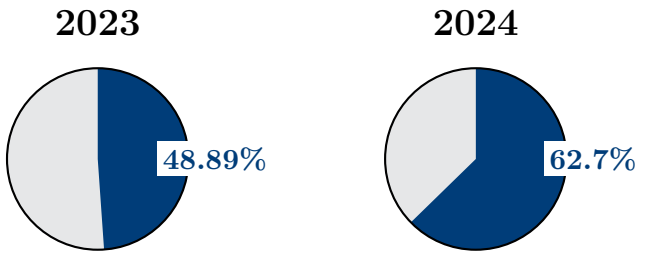

Figure 1: Use of generative AI in companies (comparison 2023 vs. 2024). Source: ISMS Forum (2024). [4]

In this context, an anonymized organization operating in a regulated and safety-critical sector identified that its traditional IT governance frameworks were insufficient for the speed and complexity of AI. The implementation timeline of the **European Union Artificial Intelligence Regulation (EU AI Act)** [1] and growing regulatory pressure across multiple jurisdictions forced a strategic rethink. The challenge was not simply to comply with the law,

but to operationalize these mandates in a way that would not stifle the innovation required to remain competitive.

2 Objective and Scope

2.1 Objective

The objective of this paper is to document and analyze an "Agile AI Governance" strategy for large organizations operating in regulated contexts. The study focuses on how governance can be operationalized without slowing innovation, through defined accountability roles, risk-based decision flows, and auditable evidence generation. The proposed model is aligned with international standards such as **ISO/IEC 42001:2023** [2] and the NIST *AI Risk Management Framework* [5].

2.2 Scope

This paper covers governance mechanisms across:

- **Solution types**: classic ML systems and GenAI/LLM-based systems.
- **Lifecycle stages**: ideation, design, development, testing, deployment, operations, and controlled retirement.
- **Control domains**: security, privacy, ethics, legal compliance, and operational resilience.
- **Operating model**: strategic, tactical, and operational governance layers, including risk classification and committee routing.

2.3 Glossary of Key Terms

To ensure a common understanding across the organization, the following key terms are defined within the governance framework:

- **CAIO (Chief AI Officer)**: The executive responsible for the overall AI strategy, ethics, and governance.
- **LLM (Large Language Model)**: Deep learning models (e.g., GPT-4) trained on vast amounts of data to understand and generate human-like text.
- **RAG (Retrieval-Augmented Generation)**: A technique that enhances LLM responses by retrieving relevant information from private enterprise data sources.
- **Prompt**: The input instruction text provided to an AI model to guide its output.
- **Hallucination**: A phenomenon where an AI generates factually incorrect or nonsensical information that appears convincing.

3 State of the Art and Regulatory Framework

AI governance has evolved from voluntary ethical guidelines to strict legal mandates. The EU AI Act establishes a risk-based approach, classifying systems into categories ranging from "unacceptable risk" (prohibited) to "minimal risk." This framework has been the organization's primary reference, given its extraterritorial reach and its influence on emerging Latin American legislation. Likewise, ISO/IEC 42001 provides the first certifiable standard for AI management

systems, offering a continuous improvement cycle (Plan-Do-Check-Act) that the organization adopted as the backbone of its operating model.

The organization's framework is anchored in several international standards and regulations [9, 31], summarized in Table 1.

Table 1: Key Regulatory and Standardization Reference Frameworks

Standard / Regulation	**Purpose and Application in the Framework**
ISO/IEC 42001	Provides the certifiable management system (AIMS) backbone and continuous improvement cycle.
EU AI Act	Establishes the risk-based legal classification and mandatory transparency requirements.
NIST AI RMF	Offers the operational taxonomy for identifying, measuring, and managing AI-specific risks.
GDPR	Ensures data protection, privacy-by-design, and individual rights in automated processing.
UNESCO Ethics	Provides the global ethical foundation for the human-centric and non-discriminatory approach.

4 Guiding Principles and Operating Frameworks

AI governance in the organization is not limited to passive compliance; rather, it is structured around **AI Governance by Design (AIGD)** principles. This approach, aligned with ISMS Forum guidelines and technical standards such as the **ISO/IEC 42001** and **ISO/IEC 23894:2023** [11] (AI risk management), requires that controls be embedded in each phase of the lifecycle:

4.1 Strategic Pillars of AI Governance

The framework is operationalized through four strategic pillars that ensure a holistic approach:

1. **Strategy and Value**: Focused on evolution, mass adoption, and ROI tracking through FinOps.
2. **Compliance and Ethics**: Managing regulatory alignment, bias detection, and fundamental rights.
3. **Security and Risks**: Implementing the AIA process, red teaming, and the AI Security Gateway.
4. **Operations and Talent**: Sustaining the AI Hub, AIOps pipelines, and the AI Champions program.

4.2 Legality and Regulatory Compliance (Compliance by Design)

All initiatives are evaluated against the requirements of Article 15 of the EU AI Act (technical robustness and cybersecurity) and the GDPR. Legal conformity is established as a non-functional requirement from the ideation stage ("Design Phase").

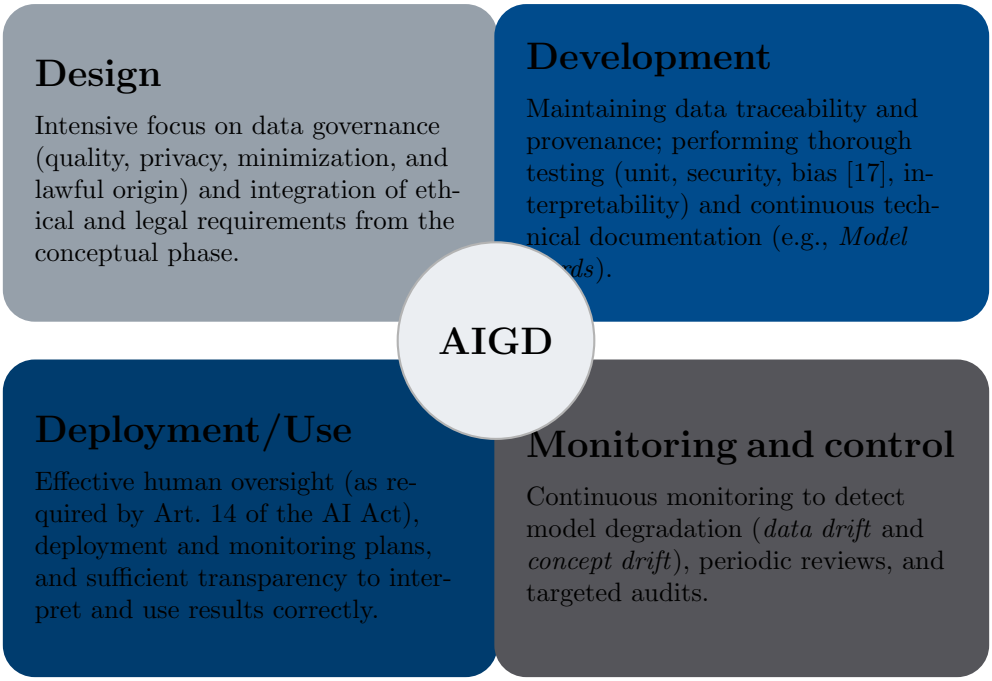

Figure 2: AI Governance by Design (*AI Governance by Design*, AIGD) across the lifecycle.

4.3 Security and Privacy (ISO/IEC 23894)

By adopting **ISO/IEC 23894**, the organization manages AI-specific risks such as *data poisoning* and *model inversion* [22, 21, 23]. Privacy is not an added layer, but an intrinsic component that uses differential-privacy techniques and data minimization at the source.

4.4 Ethics, Fairness, and Non-Discrimination (ISO/IEC TR 24368)

To mitigate bias, the organization follows the **ISO/IEC TR 24368** guidelines [12] on ethical and societal concerns. Teams conduct "Ethical Impact Assessments" (EIA) inspired by UNESCO methodology [13], validating that training data are representative and that the model does not discriminate against protected variables.

4.5 Human Oversight (Human-in-the-loop)

In line with Article 14 of the AI Act, the framework mandates that no high-risk system may operate fully autonomously. Technology acts as a copilot, keeping *accountability* with the human operator, who retains the authority to override the algorithm's decision.

5 Implemented Governance Model

The model is operationalized through organizational reengineering that defines clear roles and a precise responsibility matrix.

5.1 Definition of the AI initiatives hub

In practice, the *AI initiatives hub* was designed as a hybrid mechanism: a technology platform and an operating process. Its purpose is to remove friction between innovation and control by consolidating ideation, risk assessment, evidence management, and production operations into a single workflow.

At the process level, the hub acts as a corporate "assembly line": every use case (pilot or product) must register a business owner, a *Model Owner*, and a *Risk Owner*, and must declare data sources, vendors, and target audience (internal/customer). At the platform level, it provides standardized forms, integrates digital approvals, and automatically generates compliance artifacts aligned with ISO/IEC 42001 and the NIST AI RMF [2, 5, 6, 7].

To sustain mass adoption without creating bottlenecks, the organization defined explicit (and public) SLAs for each stage: (1) pre-check review and routing assignment in under 48 business hours, (2) risk assessment and minimum requirements within 5 business days for low/limited-risk initiatives, and (3) a weekly tactical committee decision window for high/critical risks. In practice, these times became a governance KPI: if the hub does not respond quickly, teams revert to out-of-band solutions (*Shadow AI*).

Finally, the hub standardizes the process **inputs** (business case, dataset/provenance, prompts, vendors, architecture, tests, and metrics) and its **outputs** (approval or veto, mitigation conditions, audit evidence, and production monitoring). This explicit interface definition made it possible to treat governance as an internal product, with policy versioning and iterative improvements.

To ensure adoption, the hub is supported by multi-channel communication:

- **Hub Portal**: A self-service platform containing documentation, templates, and the registry.
- **Collaboration Channel**: A dedicated internal channel for real-time tactical queries and community support.
- **AI Champions**: A network of trained individuals within business units who provide localized first-level assistance.

5.2 Initiative taxonomy: GenAI vs. classic ML

The organization explicitly separated controls for two families of solutions: (1) **classic ML** (prediction, classification, optimization) and (2) **GenAI** (conversational assistants, text/image generation, copilots, and agents). This distinction is critical because GenAI introduces specific attack vectors (e.g., *prompt injection*) and non-deterministic output risks.

Operationally, each initiative is labeled from the ideation phase by: *model type* (in-house/foundation), *mode of operation* (assisted vs. autonomous), *impacted population* (employees/customers), and *data sensitivity level*. This taxonomy feeds the risk classifier and determines whether the workflow must go through the tactical or strategic committee.

In practice, the organization defined a simple but actionable catalog of solution types (to avoid semantic debates and accelerate evaluation):

- **Internal chatbot**: employee support and access to corporate knowledge (RAG).
- **Customer service assistant**: passenger interactions and content with reputational impact.
- **Code copilot**: development assistance and IP/secret leakage risk.
- **Classification/prediction**: supervised models for operational optimization.
- **Vision/biometrics**: verification, security, and access-control use cases (typically high-risk).
- **Tool-using agents**: task automation (e.g., ticket creation, system queries, action execution).

Additionally, the *degree of autonomy* was classified into three levels: (A) recommended (suggests only), (B) assisted (executes with human confirmation), and (C) autonomous (executes without confirmation). This dimension proved decisive for setting *human-in-the-loop* requirements, evidence, and robustness testing.

The relationship between the type of solution and its primary control focus is summarized in the Solution & Control Matrix (Table 2), used here as an illustrative reference.

Table 2: Solution & Control Focus Matrix

Solution Family	**Critical Control Focus**
Code Copilots	Intellectual property protection, secret/credential leakage prevention.
Internal RAG Bots	Contextual privacy masking, factuality (hallucination) control.
Customer-Facing Bots	Brand protection, reputational filters, abuse detection.
Autonomous Agents	Human-in-the-loop validation, granular tool authorization.

5.3 Data, prompts, and audit-logging policy

To minimize information-leakage risk, the organization established GenAI-specific data-handling rules: an explicit ban on sending secrets (credentials, keys, proprietary code) and unnecessary personal data; context minimization; and separation between corporate knowledge (RAG) and conversation.

Specifically, an operational classification of information for prompts was adopted:

- **Allowed** (default): public information or non-sensitive internal information, and anonymized summaries.
- **Restricted**: PII (personal data), IP (intellectual property), and sensitive operational information; allowed only with justified need, masking controls, and an approved vendor under third-party assessment.
- **Prohibited**: secrets (keys, passwords, tokens, certificates), regulated financial data (e.g., PCI), and biometric data, except for explicitly approved high-risk cases with reinforced controls.

In addition, an audit-oriented *logging* policy was defined: which prompts/responses are recorded, for how long, under what access controls, and with what masking level. As a general rule, full prompt logs are retained only as strictly necessary (e.g., 30 days), and the organization favors recording metadata (hashes, categories, provider, prompt/model version) for audit and traceability.

In practice, this policy proved as important as the model itself, since logs often become "new datasets" that require data governance.

5.4 Controls against *prompt injection* and tool-using agents

In GenAI use cases with tool access (e.g., internal search, ticket creation, action execution), risk increases because the model can be manipulated by instructions embedded in input data (*indirect prompt injection*) or by untrusted content. Therefore, the design incorporated controls to separate instructions from data, allowlists per tool, and explicit action validation (human confirmation or deterministic rules) [10, 27, 28, 29].

In parallel, the organization institutionalized an iterative *red teaming* cycle for GenAI, using reproducible attack campaigns before production go-live and targeted reviews when prompts, models, or knowledge sources change [18, 19, 20].

5.5 Evidence model: from idea to production

To avoid "Pilot Purgatory" and sustain audits, the hub standardized a minimum evidence set structured around five specialized audit dimensions: **Technical** (robustness/security), **Functional** (performance/ROI), **Ethical** (fairness/explainability), **Organizational** (roles/governance), and **Sustainability** (carbon footprint). The core artifacts include the *AIA scorecard*, *Model Card*, *Datasheet*, robustness tests, and *fairness* reports. Practical experience showed that without these artifacts, technical and operational debt grows silently, degrading maintainability and production trust [8, 14, 15, 16].

Beyond technical metrics, the model ensures **organizational traceability** (*trazabilidad organizativa*), linking every AI decision to its legal basis, risk assessment, and human oversight record. This distinction between auditoría (the check), trazabilidad (the path), and explicabilidad (the why) is fundamental for regulatory compliance and accountability.

To achieve *end-to-end* traceability, each piece of evidence was linked to a unique initiative identifier and stored as a versioned artifact (not as standalone documents). In practice, the hub stored: model version, prompt version, dataset version (or a lineage reference), test results, and approval records. This made it possible, when a use case changed provider or a prompt was adjusted, to reconstruct which controls were re-executed and which evidence supported the go-live decision.

In initiatives involving personal data, legal evidence was integrated into this same workflow, ensuring that the DPIA (or equivalent assessment) did not become decoupled from the model lifecycle.

5.6 AI Use Case Lifecycle

The operationalization of the governance model follows a structured lifecycle that integrates multiple organizational roles through a swimlane workflow (see Figure 3). This lifecycle ensures that from the initial business idea to continuous monitoring in production, every AI initiative passes through mandatory checkpoints and evidence generation gates.

5.7 Production metrics, *drift*, and incidents

Governance extended monitoring beyond latency and availability to include response quality (e.g., hallucination rate with human sampling), *override rate*, *data drift*, and abuse signals (e.g., *jailbreak* attempts, exfiltration patterns). In case of degradation, the workflow includes *fallback* (deterministic rule or human), a "kill switch", and a *post-mortem* process with corrective actions.

To prevent monitoring from becoming an abstract exercise, the framework defines minimum thresholds by use-case class. For example, if the *override* rate exceeds a sustained threshold (illustratively, 10% weekly), or if quality sampling detects severe hallucinations above the agreed threshold, the use case enters a "controlled" mode (increased human oversight) until a technical review is completed.

Incident management was formalized with AI-specific *playbooks*: (1) severe hallucination with reputational impact, (2) suspected data leakage via prompts/responses, (3) successful *jailbreak*/prompt injection, and (4) degradation due to *drift*. Each *playbook* defines severity, containment ("kill switch", blocking RAG sources, prompt changes, disabling tools), notification to Legal/DPO when applicable, and a *post-mortem* with preventive actions [10].

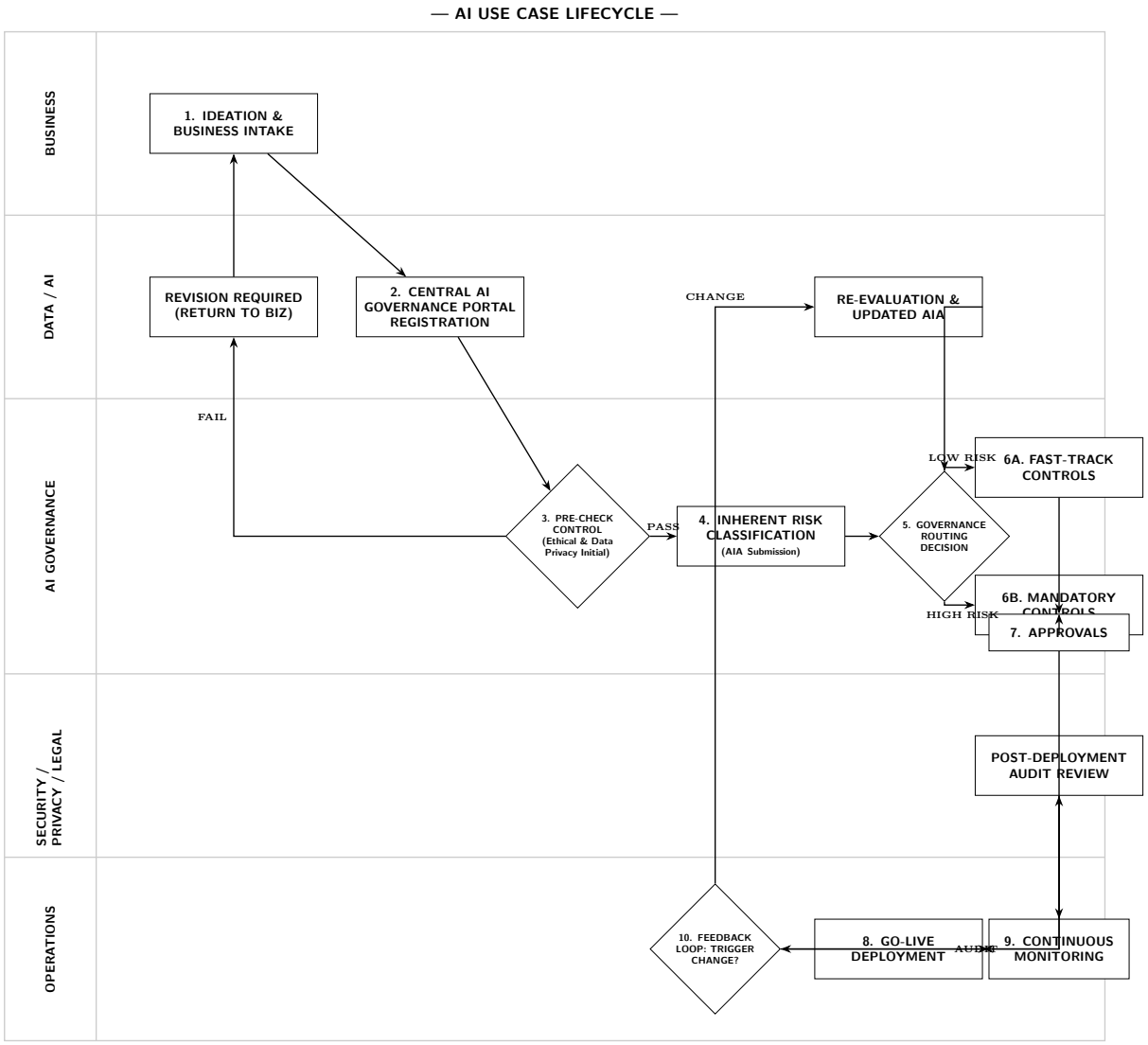

Figure 3: AI Use Case Lifecycle Diagram - Aligned with the AI governing board workflow.

Operations are supported by MLOps practices, particularly to control changes (model/prompt versioning), automate testing, and reduce regression risk when scaling use cases [24, 25, 26].

5.8 Governance dashboard, cadence, and escalation route

To make governance auditable and repeatable, the model defines a minimum dashboard and fixed review cadence:

- **Portfolio coverage**: total active initiatives and percentage registered in the central inventory.
- **Risk posture**: distribution by risk tier and percentage with required DPIA/FRIA completed.
- **Flow efficiency**: median triage time and approval cycle time by risk tier.
- **Safety and reliability**: severe hallucination rate, override rate, drift alerts, and incident count by severity.

- **Security/compliance**: open red-team findings, unresolved exceptions, and vendor due-diligence completion rate.
- **Economics**: spend variance, unit-cost trend, and risk-adjusted ROI.

Cadence is tiered by decision level: **weekly** operational follow-up, **monthly** tactical committee review, and **quarterly** strategic reporting. Escalation follows explicit thresholds:

- Any threshold breach (quality, security, privacy, or cost) triggers controlled mode within 24 hours.
- Any issue not stabilized within 5 business days escalates from operational owners to the tactical committee.
- Severity-1 incidents, regulatory exposure, or rights-impact events escalate to strategic governance and Legal/DPO within 24 hours.

5.9 Fast-track and exceptions regime

To avoid slowing innovation, a fast lane was implemented for low-risk initiatives: *fast-track* is allowed only under explicit conditions (no sensitive data, no full autonomy, no impact on rights) and with strict time limits.

Exceptions are treated as temporary risk acceptances, not permanent waivers:

- **Low/Limited-risk exceptions**: maximum validity of 90 calendar days; one renewal of up to 90 additional days.
- **High/Critical-risk exceptions**: maximum validity of 30 calendar days; weekly control reviews; renewal only with strategic approval and documented mitigation progress.
- **No indefinite exceptions**: expired exceptions automatically suspend affected capabilities until re-approval.

Minimum evidence for approval and renewal includes written business justification, affected assets/use cases, compensating controls, accountable owner, expiry date, monitoring plan, rollback/exit plan, and formal approval records. Closure requires remediation evidence, residual-risk reassessment, repository update, and explicit "closed" status in the inventory.

5.10 Standardized operational teams (AI governing board)

In the current implementation, the governance is operationalized through eleven specialized teams that constitute the *AI governing board*. These teams ensure end-to-end coverage of the AI lifecycle:

- **Audit Team**: responsible for independent sampling, evidence verification, and reporting to the audit committee.
- **Data Protect Team**: focuses on privacy-by-design, executing PII masking strategies and supporting the DPO.
- **Data Compliance Team**: monitors regulatory alignment (e.g., EU AI Act, GDPR) and coordinates mandatory impact assessments.
- **Security Team**: defines the technical security baseline, executes *red teaming*, and manages the AI gateway infrastructure.
- **GRC (Governance, Risk, and Compliance)**: orchestrates the overall framework, manages the *Ethical Dilemmas Inventory*, and monitors risk appetite.

- **CDC (Center of Cyber Defense)**: ensures the integrity of AI-generated content, coordinates human-in-the-loop validation, and manages reputational filters.
- **Data Governance Team**: accountable for data lineage, quality assurance, and lawful provenance of training/inference datasets.
- **AIOps & FMOps Team**: manages operational monitoring, scalability, and automated CI/CD/CT pipelines for models and prompts.
- **Tech Team**: responsible for the technical development, fine-tuning, and execution of AI initiatives, acting as the primary builder.
- **Legal Team**: provides contractual safeguarding, manages intellectual property rights, and handles third-party vendor legal frameworks.
- **FinOps Team**: tracks the economic impact, ROI, and cost efficiency of AI investments to prevent "Pilot Purgatory" financial leakage.

The framework also explicitly defines who approves exceptions: the tactical committee can approve low/limited-risk exceptions with compensating controls; any exception involving sensitive data, high autonomy, or customer impact requires approval at the strategic level.

5.11 Roles and responsibilities ecosystem (RACI matrix)

To operationalize governance without ambiguity, the organization defined the following responsibility assignment matrix (*Responsible, Accountable, Consulted, Informed*), detailed in Table 3.

Table 3: AI Governance RACI Matrix - Aligned with AI governing board

Activity / Task	**CAIO**	**Aud.**	**D.Prot.**	**D.Comp.**	**Arq.Cib.**	**GRC**	**CDC**	**G.Datos**	**AIOps**	**Tech**	**Legal**	**FinOps**
Strategy definition	**A**	I	I	C	C	**R**	I	C	I	I	C	C
Solution architecture	**A**	I	I	I	C	C	I	C	C	**R**	I	I
Security/GRC assessment	C	**A**	C	I	**R**	I	I	I	I	I	I	I
Compliance (AI Act)	C	I	C	**R**	I	C	I	I	I	I	**A**	I
Data governance	**A**	I	C	C	I	I	I	**R**	I	I	I	I
Model devel. & tuning	**A**	I	I	I	I	I	I	C	C	**R**	I	I
Ethical assessment (AIA)	**A**	I	I	C	I	**R**	I	I	I	I	C	I
Content integrity	C	I	I	I	I	C	**R**	I	I	I	I	I
Cost & ROI monitoring	**A**	I	I	I	I	C	I	I	I	I	I	**R**
Incident response	I	I	C	C	**R**	I	C	I	**R**	C	I	I
Evidence audit	I	**R**	I	I	I	**A**	I	I	I	I	I	I

R: Responsible, **A**: Accountable, **C**: Consulted, **I**: Informed.

Abbreviations: Aud.=Audit, D.Prot.=Data Protect, D.Comp.=Data Compliance, Arq.Cib.=Cyber Architecture, G.Datos=Data Governance.

5.12 Strategic risk management

Beyond technical risks, the model addresses critical business risks identified in the maturity analysis (based on tools such as *dataMat*):

1. **Strategic misalignment ("Pilot Purgatory")**: the risk of investing in pilots that never scale. Governance requires a clear business case before any development, ensuring alignment with corporate OKRs/KPIs. **FinOps** plays a crucial role here, providing continuous visibility into the economic impact and cloud consumption costs of each experiment.

2. **Vendor lock-in and dependency**: the risk of excessive dependency on a single LLM provider (e.g., GPT-4), which may increase costs or restrict portability. This is mitigated through an abstraction layer (GenAI Gateway) and the use of multi-vendor or open-weights models where feasible.

3. **Reputational risk**: hallucinations, disinformation, or inappropriate responses that could damage the brand. This is controlled through the **CDC** (Content Delivery & Controls) team, which implements content filters, *human-in-the-loop* oversight, and real-time response monitoring.

5.13 Third-party governance and vendor due diligence

Third-party governance is implemented through a two-layer assessment package so that sourcing, approval, and audit checks are evidence-based:

- **Use-case requester questionnaire** (business/technical owner): privacy clauses, MFA/SSO posture, log and RBAC capabilities, internal vs. customer exposure, model ownership for generated outputs, human validation points, architecture diagram, and provider/model selection rationale.

- **Vendor assurance questionnaire**: data classification and retention behavior, model retraining policy on enterprise data, SIEM/SOAR integration, anomalous-activity detection, OWASP LLM attack protections, anti-exfiltration and anti-model-extraction controls, API availability safeguards, and security operating documentation.

Procurement and production approval require auditable completion of both layers and contractual safeguards covering: data-use limitations, confidentiality, IP terms, incident-notification SLA, audit rights, subprocessor transparency, and secure deletion/return obligations at contract end.

5.14 Committee structure

Decision-making is distributed across three levels to ensure agility without losing control:

- **Strategic level (Executive Committee)**: approves high/critical-risk initiatives and defines the company's ethical risk appetite.

- **Tactical level (AI Governance Group)**: central orchestrator; reviews metrics and incidents, and manages the inventory of ethical dilemmas.

- **Operational level (Business Hubs & AIOps)**: decentralized execution; product teams own their models, supported by "AI Champions" who act as a liaison with central governance.

5.15 Operational training, culture, and awareness program

The governance model is supported by role-based enablement with mandatory completion checkpoints:

- **Business owners and product managers**: use-case framing, risk declaration quality, and human oversight obligations.
- **Engineering teams (AIOps/FMOps/Tech)**: secure LLM integration, prompt-injection defenses, evaluation pipelines, and rollback discipline.
- **Control functions (Security, Privacy, Legal, Audit)**: AI-specific control testing, evidence review, and incident playbooks.
- **Committees and executives**: risk appetite calibration, escalation criteria, and exception governance.

Operational cadence includes onboarding within 30 days of role assignment, annual recertification, quarterly simulation drills (prompt injection, leakage, and misuse scenarios), and semiannual cross-functional tabletop exercises. Program effectiveness is tracked using completion rate, assessment pass rate, incident recurrence linked to human error, and time-to-escalate during drills.

The vertical flow of accountability and support is illustrated in Figure 4 as a reference governance pattern.

Figure 4: Organizational hierarchy and governance levels.

6 Risk Management Process and Impact Assessment (AIA)

The operational core of the model is the **Algorithmic Impact Assessment (AIA)**, inspired by Canada and EU methodologies [8]. This process follows the cyclic **ISO/IEC 27005** risk management standard (Plan-Do-Check-Act) [34] and evaluates risk at two moments:

1. **Inherent risk**: the intrinsic risk of the initiative before controls are applied (e.g., use of biometric data).
2. **Residual risk**: the remaining risk once mitigation measures (technical and organizational) are implemented.

Furthermore, the model incorporates the classification of **Systemic Risk** for General-Purpose AI (GPAI) models [1], which considers factors like compute capacity (10^{25} FLOPs) and broad societal impact, requiring reinforced monitoring and documentation beyond standard high-risk categories.

6.1 Risk classifier and ethical tiering

The organization implemented an automated classification model, detailed in Table 4, which determines the level of scrutiny required for each initiative.

Table 4: AI Risk Classification Levels

Risk Level	Definition	Examples	Governance Requirement
Unacceptable	Clear threat to safety or fundamental rights.	Social scoring, Subliminal manipulation.	**Prohibited**
High	Significant impact on health, safety, or access to essential services.	Biometrics, CV filtering (HR), Credit scoring.	Conformity assessment, Human-in-the-loop, Strategic Committee.
Limited	Risk of manipulation or deception (impersonation).	Chatbots, Image/video generation.	Transparency (watermarking), Inventory registration.
Low	Minimal or no risk to individuals.	Spam filters, Predictive maintenance, Videogames.	Fast-track (self-certification).

6.2 Weighted intake questionnaire and scoring logic

To operationalize classification without exposing proprietary internal tooling names, the framework uses a weighted intake questionnaire. Each answer is mapped to a severity value and aggregated into a normalized risk score.

Each question is scored on a 0–5 scale and aggregated by dimension. The overall score is the weighted sum of dimension-level scores, normalized to a 0–100 range.

Table 5: Weighted Intake Dimensions for Risk Routing

Dimension	Weight	Representative Signals
Data sensitivity and rights impact	25%	PII/PCI/biometrics, potential impact on customer or employee rights.
Exposure and affected population	20%	Internal-only vs. customer-facing use, expected user scale.
Autonomy and tool execution	15%	Suggestion-only behavior vs. autonomous actions/tool-calling.
Model and solution type	15%	Classic ML vs. LLM/GenAI, fine-tuning needs, architecture complexity.
Security integration posture	10%	Approved gateway integration, IAM/SSO, logging and monitoring readiness.
Third-party dependency	10%	External API/model dependency, due-diligence and contractual posture.
Regulatory/geographic complexity	5%	Cross-border processing and multi-jurisdiction obligations.

The score is mapped to operating tiers: **Low** (0–24), **Limited** (25–49), **High** (50–74), and **Critical** (75–100). Hard gating rules override numeric scoring:

- Legally prohibited use patterns are classified as **Unacceptable** and blocked.

- Customer-facing autonomous actions involving sensitive data are escalated to strategic approval.
- External LLM use outside the approved gateway path triggers mandatory remediation before go-live.
- Missing legal/vendor safeguards blocks procurement and production deployment until closure.

6.3 Multidimensional audit and ethical KPIs

Pre-deployment validation is deepened through an audit that interrogates the system across five dimensions, feeding a corporate **Ethical Dilemmas Inventory**:

6.3.1 1. Technical dimension (robustness)

- Is the model resilient to adversarial machine learning attacks?
- Is there a *fallback* mechanism (deterministic rule or human) if the model fails or drops below a confidence threshold?

6.3.2 2. Functional dimension (fitness for purpose)

- Does AI solve the problem better than a traditional heuristic?
- Is there a clear business success metric (e.g., 15% reduction in false positives)?

6.3.3 3. Ethical dimension (fairness)

- Has statistical parity been measured for different demographic groups?
- Are **ethical impact KPIs** defined to monitor potential harm in production?

6.3.4 4. Organizational dimension (governance)

- Is the *Model Owner* clearly identified and trained?
- Has data lineage been documented from source to inference?

6.3.5 5. Sustainability dimension (Green AI)

- Have CO_2 emissions associated with model training and operation been estimated?
- Is the energy cost justified by the business value generated?

6.4 Audit checklist

To ensure consistency across reviews, the internal audit team uses a standardized checklist (Table 6) that must be completed before go-live.

7 Data Protection, Privacy, and Bias Prevention

The convergence of the GDPR and the EU AI Act establishes a dual mandate for organizational transparency. While the GDPR focuses on the individual's right to understand automated processing (Art. 22), the AI Act mandates technical documentation and systemic transparency (Art. 50) to ensure trustworthiness [1, 35].

Table 6: Audit Controls Checklist by Dimension

Dimension	Control / Audit Question	Evidence
Technical	Has an AI-specific penetration test (*Red Teaming*) been performed? [18, 19]	Pentest report
	Is there a deterministic *fallback* mechanism in case of model failure?	Architecture doc
	Has robustness against *data poisoning* been validated? [22]	N/A or test report
Functional	Does the model outperform the baseline (human or heuristic)?	Confusion matrix
	Are clear business KPIs defined (e.g., % churn reduction)?	Business case
Ethical	Has the Algorithmic Impact Assessment (AIA) been executed?	AIA scorecard
	Has statistical parity been verified for protected groups?	Fairness report
	Is the decision explainable for a non-technical end user?	Explainability guide
Legal / Org	Is the *Model Owner* appointment signed?	Committee minutes
	Has a DPIA been performed if personal data are involved?	Approved DPIA
Sustainability	Has the carbon footprint of training and inference been estimated?	CO_2 calculator

7.1 Data minimization and DPIA/FRIA integration

The proposed model applies the principle of **data minimization** (Art. 5.1c GDPR) not only during operation but also during the training and fine-tuning phases. This involves the use of pseudonymization, differential privacy, and synthetic data generation where possible to reduce the surface of personal data exposure.

A critical operational innovation is the integration of the **Data Protection Impact Assessment (DPIA)** and the **Fundamental Rights Impact Assessment (FRIA)**. While they remain distinct legal instruments, their execution is synchronized to identify synergies:

- **GDPR DPIA**: evaluates risks to personal data privacy.
- **AI Act FRIA**: evaluates risks to non-discrimination, freedom of expression, and human dignity (required for high-risk systems under Art. 27).

7.2 Managing the bias lifecycle

To prevent discriminatory outcomes, the governance framework identifies and mitigates bias across three levels:

1. **Dataset bias**: ensuring representativeness and quality in the training sets to avoid historical prejudices.
2. **Personal bias**: mitigating the influence of developer-level subjective perceptions during model tuning.
3. **Algorithmic bias**: auditing the model logic to ensure it does not prioritize variables that correlate with protected categories (e.g., zip code correlating with ethnicity).

8 Ethics, Intellectual Property, and Emerging Challenges

The rise of General-Purpose AI (GPAI) and agentic systems introduces ethical dilemmas that transcend traditional risk management. The organization has adopted a "Human-Centric AI" approach, prioritizing cognitive autonomy and transparency.

8.1 Intellectual property and the GPAI paradox

The ease of use of foundation models masks the complexity of their training data provenance. We have implemented a "Copyright Clearance" check within the AI initiatives hub to evaluate the legitimacy of datasets and address the risks associated with using copyrighted material without authorization—a focus of ongoing litigation such as *New York Times v. OpenAI*.

Furthermore, the legal status of AI-generated content remains an open question: who holds the authorship? Our policy establishes that AI is a tool, and responsibility (as well as IP exploitation rights where legally permissible) remains with the human operator and the sponsoring business unit.

8.2 Epistemic trust and deepfakes

In an era of hyper-realistic synthetic media, maintaining "epistemic trust" is paramount. The framework mandates the labeling of all AI-generated content (e.g., watermarking or disclaimer tags) to prevent deception, following the transparency requirements for "Limited Risk" systems.

8.3 Future outlook: Neuro-rights and cognitive autonomy

Looking ahead, we recognize emerging challenges such as neuro-rights: the protection of mental integrity against neuro-technologies and highly persuasive AI agents. Inspired by Chile's 2021 constitutional reform [33] and UNESCO recommendations [30], our ethics committee proactively reviews use cases that could impact subliminal perception or cognitive freedom, ensuring AI remains a partner rather than a manipulator.

Governance is supported by a centralized technology platform, the **AI initiatives hub**, which operationalizes the controls defined on paper:

- **GenAI Gateway (Cognitive Firewall)**: acts as the single gateway for consuming external LLMs. It applies real-time "data hygiene" rules [10], managed by the **Cybersecurity Architecture** and **CDC** teams:
  - *On-the-fly anonymization*: detects and masks personal data (PII) using regular expressions and NLP before sending the prompt to the provider.
  - *Secrets filter*: blocks sending API keys, passwords, or proprietary code.
- **Observability module**: continuous monitoring of technical metrics (latency, throughput) and business metrics by the **AIOps** and **FinOps** teams. Detects *data drift* and cost anomalies that could degrade model performance or ROI.
- **Asset and vendor registry**: a centralized inventory linking each model to its owner, risk tier, and "life sheet" (Model Card) [14, 15], ensuring traceability required by ISO/IEC 42001 and maintained by **GRC**.

9 Lessons Learned (2025) and 2026 Outlook

Across pilot applications and benchmarked programs, culture consistently appears as the most critical vector. Recent industry studies support this observation: according to the *Cisco AI*

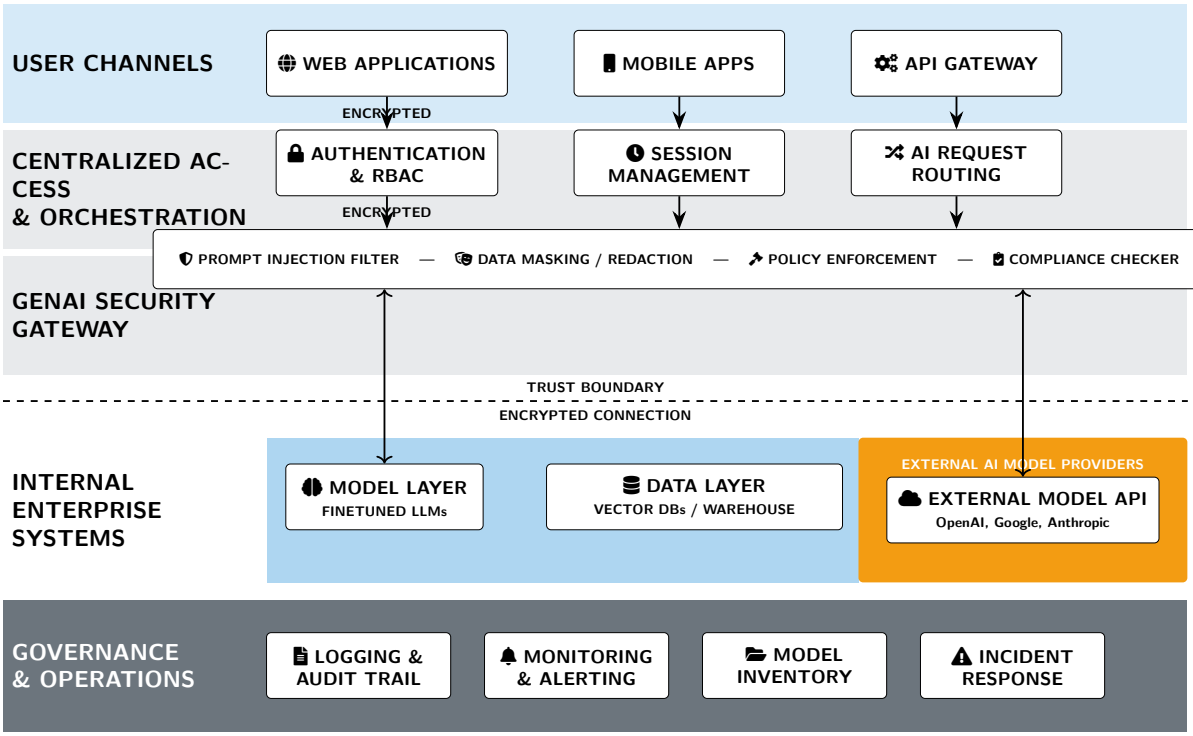

Figure 5: Reference security architecture for GenAI.

Readiness Index 2025 [32], only 23% of organizations have their governance processes fully prepared for the challenges of AI. Furthermore, 93% of organizations with advanced AI adoption have already integrated specific AI risks into their data protection policies, highlighting that maturity in governance is strongly correlated with privacy rigor.

However, a significant gap remains: only 34% of companies trust their current cybersecurity infrastructure to face AI-specific threats such as prompt injection or poisoning. In illustrative implementations, the risk of "Shadow AI" is often mitigated more effectively by offering secure corporate alternatives (AI initiatives hub) and fostering "AI risk literacy" than by imposing draconian prohibitions.

For 2026 and beyond, the outlook in comparable programs is to pursue **ISO/IEC 42001** readiness, automate audit-evidence collection via APIs, and extend the "AI Security Champions" program across business units to strengthen the human firewall.

The roadmap is parameter-driven: depending on organizational risk and maturity indicators, specific strategic milestones become necessary, as detailed in Table 7.

10 Evaluation Method and Observed Results

10.1 Evaluation design

The framework was evaluated through a pre/post operational analysis over a 12-month implementation window, using governance workflow logs, inventory records, audit artifacts, and incident tickets. The baseline corresponds to the period before centralized governance controls were enforced; the post period corresponds to the first year after rollout.

The goal was not to establish strict causal inference, but to measure whether governance operating discipline improved key risk, speed, and control-completeness indicators.

Table 7: Parameter-Based Agile AI Governance Milestones

Activation Parameter	Strategic Milestones
Portfolio Scale	When the active AI portfolio grows beyond pilot scale (e.g., 20+ initiatives or multi-unit demand), establish a centralized governance hub, baseline prompt-injection controls, and mandatory pre-check routing.
Regulatory/Data Exposure	When use cases involve personal data, sensitive business information, or cross-border processing, deploy full asset inventory traceability, automated AIA scoring, and registration of ethical/legal dilemmas.
Residual Risk Profile	When residual risk remains high/critical after controls or customer-facing impact increases, require independent external assurance, reinforced evidence controls, and certification-readiness activities aligned with **ISO/IEC 42001**.
Autonomy and Agentic Execution	When autonomous or tool-using agents operate at enterprise scale, implement federated governance, stricter human-in-the-loop gates, and AI-specific incident playbooks with continuous re-evaluation loops.

10.2 Observed outcomes

Table 8 summarizes the primary indicators tracked in the implementation window.

Table 8: Observed Governance Outcomes (Pre vs. Post Implementation)

Indicator	Baseline	Post	Change
Inventory coverage (registered initiatives / estimated total)	41%	96%	+55 pp
Median triage time (business days)	8.5	2.0	-76%
Median approval cycle for low/limited risk (business days)	18.0	6.0	-67%
Initiatives with complete evidence package at go-live	27%	89%	+62 pp
Severity-1 AI incidents per quarter	3	1	-67%
Mean time to contain AI incidents (hours)	19	5	-74%
Vendors with complete due diligence before go-live	48%	94%	+46 pp

These results indicate simultaneous progress in governance velocity and control rigor. In particular, the reduction in triage/approval times suggests that governance can operate as an enablement layer rather than a bottleneck when process routing, evidence templates, and risk thresholds are standardized.

11 Limitations and Threats to Validity

Despite encouraging outcomes, the case study has constraints:

- **Single organizational context**: results may not transfer directly to organizations with different risk appetite, tooling maturity, or operating model.
- **Pre/post observational design**: confounding factors (team growth, parallel tooling improvements, process learning) may influence measured improvements.

- **Data-quality dependency**: indicator quality is bounded by the completeness and consistency of operational logging.
- **Regulatory evolution**: some control mappings may require updates as AI-specific regulation evolves across jurisdictions.

12 Control-to-Regulation Traceability Matrix

To support audit-readiness, core controls are mapped to applicable normative expectations (Table 9).

Table 9: Control-to-Regulation Traceability Matrix

Internal Control	**Normative Anchor**	**Evidence Artifact**
Risk-based intake and tier routing	EU AI Act risk-based approach; NIST AI RMF governance functions	Risk scorecard and routing record
Human-in-the-loop for high-impact use cases	EU AI Act human oversight requirements	Approval conditions and override logs
Security testing and robustness checks	EU AI Act robustness/cybersecurity; ISO/IEC 23894 risk guidance	Red-team report and test evidence
Privacy impact and minimization controls	GDPR data minimization and DPIA obligations	DPIA/FRIA package and data map
Model inventory and lifecycle evidence	ISO/IEC 42001 management-system traceability	Model card, datasheet, version history
Third-party due diligence and contracts	ISO-aligned supplier governance and security assurance practices	Vendor questionnaire and contract clauses

13 Framework Review and Validity

This framework is reviewed at least annually, and earlier whenever a material trigger occurs: relevant regulatory changes, high-severity incidents, major architecture shifts (e.g., expanded agentic autonomy), or substantial changes in third-party dependency.

The governance office coordinates updates with Security, Privacy, Legal, Audit, and Engineering stakeholders. Each release is versioned and published with effective date, transition requirements for in-flight initiatives, and controls that must be revalidated.

14 Operational Annexes (Published Extract)

The complete annexes are maintained as controlled documents. This paper includes a publication-safe extract of their minimum structure.

14.1 Annex A: Initiative intake template (minimum fields)

14.2 Annex B: Weighted questionnaire extract

14.3 Annex C: BPMN flow gates (summary)

1. Ideation and business intake.

Field	Minimum content
Initiative identity	Unique ID, sponsor, business owner, model owner, risk owner
Business definition	Problem statement, KPI target, expected users, internal vs. external exposure
Data profile	Data categories (PII/PCI/etc.), origin, residency constraints, retention expectations
Technical profile	Solution type (ML/GenAI), architecture summary, provider/vendor footprint
Control profile	Required assessments, human oversight mode, fallback and rollback strategy

Category	Example question	Scale
Data sensitivity	Does the use case process personal, financial, or biometric data?	0–5
Exposure	Is the solution customer-facing or internal-only?	0–5
Model behavior	Is the model autonomous or tool-executing without confirmation?	0–5
Security posture	Is the solution integrated with approved gateway, IAM, and logging?	0–5
Vendor dependency	Is there external API/provider dependency with complete due diligence?	0–5

2. Central registration and initial quality checks.
3. Risk scoring and route assignment.
4. Specialized reviews (security, privacy, legal, architecture).
5. Approval decision (tactical or strategic).
6. Controlled deployment with rollback readiness.
7. Production monitoring and incident handling.
8. Change/reassessment or controlled retirement.

14.4 Annex D: Third-party due-diligence package (extract)

- **Requester checklist**: contractual privacy clauses, MFA/SSO readiness, logs, RBAC, architecture, and human validation path.
- **Vendor checklist**: data retention policy, model retraining policy, OWASP LLM protections, SIEM integration, exfiltration controls, and incident response SLA.
- **Required artifacts**: completed questionnaires, security/privacy assessments, and signed contractual safeguards.

15 Expected Benefits

The implementation of the agile AI governance model and the centralized hub is expected to deliver a wide range of strategic, operational, and technical advantages for the organization:

- **Risk Mitigation**: Significant reduction of security and compliance risks across the AI portfolio.
- **Operational Efficiency**: Increased efficiency through systematic reuse of approved prompts and technical components.
- **Ecosystem Traceability**: Total traceability of the enterprise agent ecosystem and decision-making history.
- **Development Agility**: Accelerated development cycles under controlled environments and clear guidelines.
- **Responsible Culture**: Promotion of a responsible and federated AI culture throughout the organization.
- **Regulatory Readiness**: Alignment with global regulations (GDPR, EU AI Act) with reduced audit friction.
- **Certification Readiness**: Clear path to international certifications (e.g., ISO/IEC 42001, ISO 27001).
- **Resilient Incident Response**: Lower Mean Time to Repair (MTTR) through AI-specific playbooks and rollback capabilities.
- **Elimination of Shadow AI**: Centralized registration of all initiatives within approved governance platforms.
- **Supply Chain Governance**: Total control over the AI supply chain through automated Model Cards and SBOMs.
- **Information Leakage Prevention**: Reinforced security via DLP, isolated execution, and secret management.
- **Business Continuity**: Resilience through multi-vendor/multi-cloud strategies and automated failover plans.
- **Advanced Observability**: Real-time monitoring of AI SLOs (quality, latency, cost, drift, and toxicity).
- **Streamlined Onboarding**: Accelerated time-to-approve and time-to-use for new AI initiatives.
- **Vendor Independence**: Reduced lock-in through support for open protocols (AGORA, MCP, A2A, ACP, AGP).
- **High-Quality Output**: Enhanced quality through the "AI Factory" approach with automated validation.
- **Privacy by Design**: Rigorous data minimization and reduction of unnecessary data sprawl.
- **Data Sovereignty**: Governed lineage, retention, and residence policies for all AI-associated data.

- **Brand Protection**: Provenance tracking and content authenticity verification (e.g., C2PA).
- **Trust and Reputation**: Greater confidence from customers and regulators in the company's AI systems.
- **Risk-Adjusted Prioritization**: Governance dashboards enabling prioritization based on both ROI and risk scores.
- **Homogeneous Upskilling**: Standardized runbooks, prompt repositories, and toolkits for all teams.
- **Federated Autonomy**: Scaling through a federated model with common standard guardrails.
- **AI FinOps**: Cost transparency through showback/chargeback and efficient autoscaling.
- **Sustainability (Green AI)**: Integration of sustainability metrics (CO_2e) for ESG reporting.

16 Conclusion

This case study shows that Agile AI Governance is not a bureaucratic brake, but a strategic enabler. By integrating *Governance by Design* principles, clear roles, and automation, the organization turned regulatory compliance from a burden into a competitive advantage grounded in digital trust. Adopting standards such as ISO/IEC 42001 and the NIST framework not only reduced legal exposure, but also accelerated the organization's ability to innovate sustainably in an uncertain regulatory environment.

References

[1] European Union, "Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)," *Official Journal of the European Union*, 2024.

[2] ISO/IEC, "ISO/IEC 42001:2023 Information technology — Artificial intelligence — Management system," *International Organization for Standardization*, 2023.

[3] ISMS Forum and Cisco, "AI Governance Study: A practical implementation guide," *Spanish Association for the Promotion of Information Security*, November 2025.

[4] ISMS Forum, "Analysis II survey: Adoption and Governance of Artificial Intelligence," *Spanish Association for the Promotion of Information Security*, 2024.

[5] NIST, "Artificial Intelligence Risk Management Framework (AI RMF 1.0)," *National Institute of Standards and Technology*, U.S. Department of Commerce, 2023.

[6] NIST, "Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile," *National Institute of Standards and Technology*, 2024.

[7] NIST, "Roadmap for the NIST Artificial Intelligence Risk Management Framework (AI RMF)," *National Institute of Standards and Technology*, 2023.

[8] Government of Canada, "Algorithmic Impact Assessment (AIA)," *Directive on Automated Decision-Making*, 2019 (online updates).

[9] European Commission (AI HLEG), "ALTAI: Assessment List for Trustworthy Artificial Intelligence," 2020.

[10] OWASP, "Top 10 for Large Language Model Applications," OWASP Foundation, 2024–2025.

[11] ISO/IEC, "ISO/IEC 23894:2023 Information technology — Artificial intelligence — Guidance on risk management," *International Organization for Standardization*, 2023.

[12] ISO/IEC, "ISO/IEC TR 24368:2022 Information technology — Artificial intelligence — Overview of ethical and societal concerns," *International Organization for Standardization*, 2022.

[13] IEEE, "IEEE 7000-2021 — IEEE Standard Model Process for Addressing Ethical Concerns during System Design," *IEEE Standards Association*, 2021.

[14] M. Mitchell et al., "Model Cards for Model Reporting," *Proceedings of FAT**, 2019 (arXiv:1810.03993).

[15] T. Gebru et al., "Datasheets for Datasets," 2018 (arXiv:1803.09010).

[16] D. Sculley et al., "Hidden Technical Debt in Machine Learning Systems," *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.

[17] P. Saleiro et al., "Aequitas: A Bias and Fairness Audit Toolkit," 2018.

[18] Y. Ge et al., "MART: Improving LLM Safety with Multi-round Automatic Red-Teaming," 2023 (arXiv:2311.07689).

[19] Y. Zhou et al., "AutoRedTeamer: Autonomous Red Teaming with Lifelong Attack Integration," 2025 (arXiv:2503.15754).

[20] H. Zhang et al., "Automated Red-Teaming Framework for LLM Security Assessment," 2025 (arXiv:2512.20677).

[21] Q. Zhang et al., "The Secret Revealer: Generative Model-Inversion Attacks Against Deep Neural Networks," 2019 (arXiv:1911.07135).

[22] J. Steinhardt, P. W. Koh, and P. Liang, "Certified Defenses for Data Poisoning Attacks," 2017 (arXiv:1706.03691).

[23] M. Jagielski et al., "High Accuracy and High Fidelity Extraction of Neural Networks," 2019 (arXiv:1909.01838).

[24] M. John, H. Holmström Olsson, and J. Bosch, "Towards MLOps: A Framework and Maturity Model," *SEAA*, 2021.

[25] "An empirical guide to MLOps adoption: Framework, maturity model and taxonomy," *Information and Software Technology*, 2025.

[26] "MLOps best practices, challenges and maturity models: A systematic literature review," *Information and Software Technology*, 2025.

[27] J. Yi et al., "Jailbreak, Prompt Injection, and Prompt Leaking Attacks against LLMs," *Microsoft Research*, 2023 (arXiv:2310.12815).

[28] S. Yao et al., "ReAct: Synergizing Reasoning and Acting in Language Models," 2022 (arXiv:2210.03629).

[29] T. Schick et al., "Toolformer: Language Models Can Teach Themselves to Use Tools," 2023 (arXiv:2302.04761).

[30] UNESCO, "Recommendation on the Ethics of Artificial Intelligence," *United Nations Educational, Scientific and Cultural Organization*, 2021.

[31] OECD, "Recommendation of the Council on Artificial Intelligence," *Organisation for Economic Co-operation and Development*, 2019.

[32] Cisco, "AI Readiness Index 2025: From Aspiration to Adoption," *Cisco Systems Inc.*, January 2025.

[33] Republic of Chile, "Constitutional Reform on Neurorights and Mental Integrity (Law No. 21.383)," *Official Journal of Chile*, 2021.

[34] ISO/IEC, "ISO/IEC 27005:2022 Information security, cybersecurity and privacy protection — Guidance on managing information security risks," *International Organization for Standardization*, 2022.

[35] AEPD, "Adecuación al RGPD de tratamientos que incorporan Inteligencia Artificial," *Agencia Española de Protección de Datos*, 2020.