



Enterprise AI at Scale: A Maturity and Deployment Research Paper for Generative, Agentic and Traditional AI

Richard Dickerson · Cognostrix

Version 8.1 — Distributed Baseline — 26 September 2026

Status: For information. Not a standard, certification scheme, legal opinion or completeness claim.

Scope, limitations and notices

Licence. This work is published by Cognostrix under a Creative Commons Attribution 4.0 International licence (CC BY 4.0). You may copy, distribute, adapt and build on it, including commercially, provided Cognostrix is credited. Suggested attribution: *Cognostrix, "Enterprise AI at Scale: A Maturity and Deployment Research Paper for Generative, Agentic and Traditional AI", v8.1, September 2026.*

No claim of originality. This research paper is a consolidation. It is drawn from published standards, regulations, supervisory guidance, vendor documentation and practitioner literature, and is credited to their sources throughout. Cognostrix claims no authorship of the underlying ideas and asserts no rights over them. Where an idea has no single identifiable source, that is a statement about the published record, not a claim of invention.

Provided as is. This document is provided without warranty of any kind, express or implied. Cognostrix accepts no liability for any loss or damage arising from its use or from reliance on it. Nothing here creates a professional or advisory relationship.

Third-party names. Standards, frameworks, regulations, products and organisations are named for identification and mapping only. All trade marks are the property of their respective owners.

Scope. This research paper focuses on mid-size and large enterprises and institutions. Its legal treatment is EU/Sweden-primary; UK, US and other-jurisdiction references are contextual pointers rather than substitutes. The operating-model, engineering, assurance, security, service-management and value principles are intended to be portable across jurisdictions.

For Information. This research paper is provided for information and practical implementation planning. It does not establish that any particular AI system, programme, organisation or control environment is safe, secure, compliant, certified or fit for purpose. It is not legal, regulatory, compliance, audit, investment or certification advice, and it does not state definitive minimum requirements for any law, standard, regulator or jurisdiction.

Applicability: Applicability, adequacy and operating effectiveness depend on the specific use case, implementation, organisation, jurisdiction, threat model, risk appetite and evidence. Authoritative legal and regulatory requirements, licensed standards and qualified specialists should be used where formal compliance, assurance or certification matters.

Endorsement. Third-party standards, frameworks, products and organisations are named for identification, comparison and source mapping only. Their inclusion does not imply endorsement of this paper by, or affiliation with, the bodies named.

Executive synthesis

The central finding of this research is that a large organization should **not treat generative AI as another software rollout, nor agentic AI as an extension of chatbot deployment**. The more durable model is to establish an **enterprise AI management and delivery capability** that combines strategy, portfolio management, governance, data, technology, security, service management, workforce transformation and measurable value realization.

That conclusion is consistent with the direction of the major governance frameworks. ISO/IEC 42001 treats AI as an organization-wide management system that must be established, maintained and continually improved; NIST's AI Risk Management Framework similarly organizes AI risk management around **Govern, Map, Measure and Manage**, with a separate Generative AI Profile covering risks specific to GenAI. [1, 9, 2]

It also reflects the practical problem visible in current enterprise research: AI usage and individual productivity gains have expanded faster than enterprise-level value realization. McKinsey's 2025 survey reported AI use in at least one business function at 88% of responding organizations while only 7% reported full enterprise scaling **under its 2025 scaling definition**; its August 2026 survey found that 80% of respondents said AI improved their individual productivity, yet only 37% attributed any EBIT impact to organizational AI use and only 6% qualified as high performers. The same research stream argues that the larger economic opportunity lies in redesigning end-to-end workflows rather than simply deploying horizontal copilots. [17, 63, 4]

The AI-related roles identified and included in this paper strongly reinforce this picture. Taken together, they do not describe an organization that merely needs "AI projects." They describe the constituent capabilities of an enterprise AI operating model: AI strategy and business cases; enterprise enablement; formal governance; EU AI Act compliance; security and risk; AI platforms; agentic architecture; use-case portfolios; deployment; SaaS intake; data classification; identity and permissions; IT service introduction; CMDB and support handover; AI literacy and training; change management; product management; production AI/ML engineering; executive communication; benefits realization; and safe operationalization.

How to use this paper

The paper is designed to be of practical use, providing a broad guide to the deployment of AI within mid-to-large scale organisations, from day zero through to the creation of a mature AI organization.

- **Executive orientation:** use the maturity journey, capability map and maturity matrix to understand the target state and the sequence of enterprise change.
- **Maturity assessment:** use the twelve capability dimensions to assess the current state, target state and evidence required to move between stages.
- **Implementation planning:** use the lifecycle and detailed capability sections to build a programme roadmap, ownership model, workstreams and gates.
- **Deep reference:** use Part V and the appendices to examine a particular domain such as security, service management, workforce, data, assurance or FinOps.
- **Machine-assisted use:** the paper can be provided to an AI/agent together with organisational context to support gap analysis, roadmap design and project planning. Human accountability remains with the organisation, and the paper must not be treated by an agent as a compliance determination.

Contents

Scope, limitations and notices	2
Executive synthesis	3
How to use this paper	3
PART I — PURPOSE AND CORE THESIS	6
1.1 Purpose of the paper	6
1.2 Core thesis	6
1.3 Extend what already exists before creating AI-specific machinery	6
PART II — THE ENTERPRISE AI MATURITY JOURNEY	7
Enterprise AI maturity journey: from first controls to pervasive AI	7
Mobilize and baseline	8
Controlled enablement	8
Governed pilots and standards co-development	9
Industrialise and build federated capability	9
Workflow and workforce rewiring	10
Adaptive AI enterprise	10
PART III — THE CAPABILITY MODEL: WHAT MUST MATURE TOGETHER	12
Target-state operating model, roles and technology foundation	12
3.1 Twelve capability domains	12
PART IV — MATURITY ACROSS THE CAPABILITY DOMAINS	14
A practical maturity model	14
The executive scorecard	16
PART V — DETAILED GUIDANCE BY CAPABILITY DOMAIN	19
5.1 Strategy and value	19
Business cases, use-case selection and portfolio management	19
5.2 Governance and accountability	24
Governance, policy, security and service introduction	24
The organizational model	25
Scale the operating model proportionately	27
Operate governance as an AI management system	28
The required policy and control library	29
Maintain a live regulatory obligations calendar	31
5.3 Security and resilience	35
Agent security and non-human identity	35
Agent identity, delegation and tool/protocol supply-chain security	35
UI-acting and computer-use agents	38
5.4 Platform and architecture	38
The technical control plane	38
5.5 Data and knowledge	40
Privacy and lifecycle controls for RAG and vector systems	40
Govern agent memory and persistent state as enterprise data	41

5.6 Delivery and engineering	42
Risk-tier the lifecycle	42
Make permissions and scope changes governance events.....	43
5.7 Systemic assurance and human oversight	43
Introduce an explicit agent autonomy ladder	43
Preserve mature model-risk management for predictive and traditional ML	44
Assure multi-agent systems as systems, not collections of individually tested agents	45
Map traditional assurance stages onto AI and agentic releases	46
Calibrate evaluation thresholds rather than inventing pass marks	47
Choose human oversight by risk and operating speed	48
5.8 Operations and service management	48
Integrate AI with IT service management	48
5.9 Vendor and third-party management	50
Treat AI procurement and third-party dependencies as lifecycle controls	50
Treat sustainability and compute capacity as proportionate architecture/procurement concerns.....	51
5.10 Economics, funding and AI FinOps	52
Make AI FinOps a runtime control, not a monthly report.....	52
5.11 People, roles and capability building	53
Roles required in the mature organization	53
Workforce enablement, training and organizational change	54
5.12 Workforce, employment and organisational change	56
Treat employment relations and workforce governance as part of deployment	56
Redesign jobs, not just tasks	57
PART VI — CONCLUSION	58
PART VII — APPENDICES AND PRACTITIONER TOOLKIT.....	59
Appendix A — Lift-and-adapt practitioner artefacts.....	59
Lift-and-adapt practitioner artefacts	59
Appendix B — Machine-assisted use of this paper	61
Appendix C — Research method and assurance boundary.....	61
Review methodology and residual assurance boundary.....	61
Appendix D — Implementation anti-patterns	62
Implementation anti-patterns to avoid.....	62
Appendix E — Related Cognostrix research	62
References and source trail.....	63
Public sources	63
Evidence interpretation note	68

PART I — PURPOSE AND CORE THESIS

1.1 Purpose of the paper

This paper is an umbrella research and implementation reference for organisations introducing, scaling and maturing generative, agentic and traditional AI. Its purpose is to make the overall enterprise system visible: the journey an organisation follows, the capabilities that must mature together, the evidence that maturity should produce, and the lower-level practices available within each capability domain.

It is deliberately broader than a governance framework, technical architecture or transformation plan in isolation. Those are constituent parts of the same enterprise capability. The paper is intended to provide the spine from which deeper domain papers, standards, procedures, templates and implementation playbooks can be referenced.

1.2 Core thesis

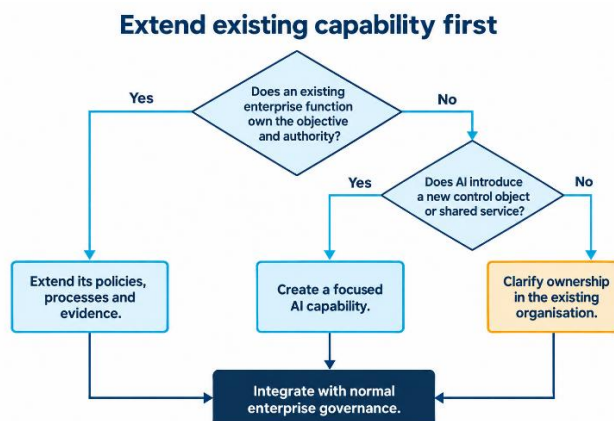
- Enterprise AI should be managed as an enduring organisational capability, not a collection of disconnected tools or pilots.
- The unit of value is normally the business workflow or outcome, not the model, chatbot or agent.
- Security, identity and risk boundaries are continuous foundations across the lifecycle, not downstream approval activities.
- Maturity is evidenced by repeatability, control, value, operating capability and delegated competence — not by the number of AI users or experiments.
- The mature target is AI-pervasive but not AI-maximalist: AI is considered as a normal design option and used where it is the appropriate mechanism.

1.3 Extend what already exists before creating AI-specific machinery

Most capabilities required for enterprise AI already exist in a mature organisation: strategy, budgeting, enterprise architecture, IAM, cybersecurity, data governance, model risk, SDLC, procurement, third-party risk, ITSM, finance, HR, legal/compliance and Internal Audit. The default should therefore be to extend existing structures, policies and decision rights with AI-specific triggers and evidence rather than rename or duplicate the enterprise around AI.

Create a dedicated AI capability where the control object or service is genuinely new, where consistency across the enterprise cannot be achieved through an existing function, or where the organisation needs a reusable shared product. Examples include an enterprise AI inventory/AIMS coordination capability, governed model access and routing, agent identity/delegation, AI evaluation infrastructure, RAG and persistent-memory controls, model/tool version governance and agent autonomy boundaries.

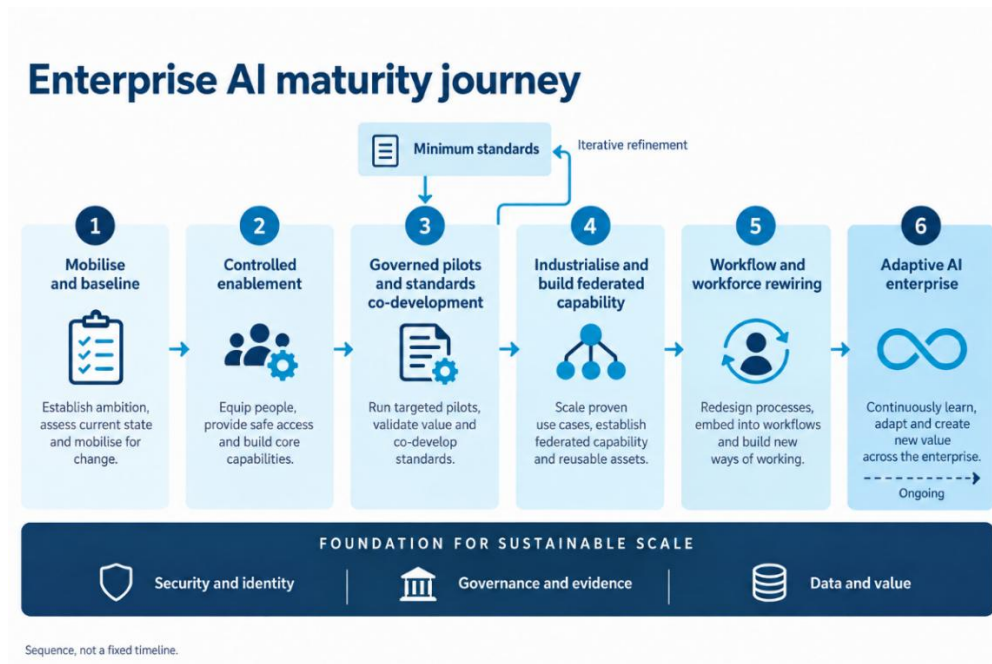
Where a mature model-risk framework already exists, it should remain the anchor for predictive and statistical ML and can often provide an important home or challenge function for broader AI governance. GenAI and agentic controls should extend — not weaken or bypass — that established discipline. The precise organisational home should follow the enterprise's existing accountability model.



PART II — THE ENTERPRISE AI MATURITY JOURNEY

The journey below is a sequence, not a fixed calendar. Organisations will progress at different speeds, and different business domains may reach different levels of delegated capability at different times. Minimum governance, security, data, legal/privacy and evaluation guardrails should exist before material pilots are approved; those guardrails are then refined through the pilot programme rather than designed in isolation from it.

Security is deliberately not shown as a one-time stage. It is a continuous constraint from first access through delegated autonomy, production operation and retirement.



Enterprise AI maturity journey: from first controls to pervasive AI

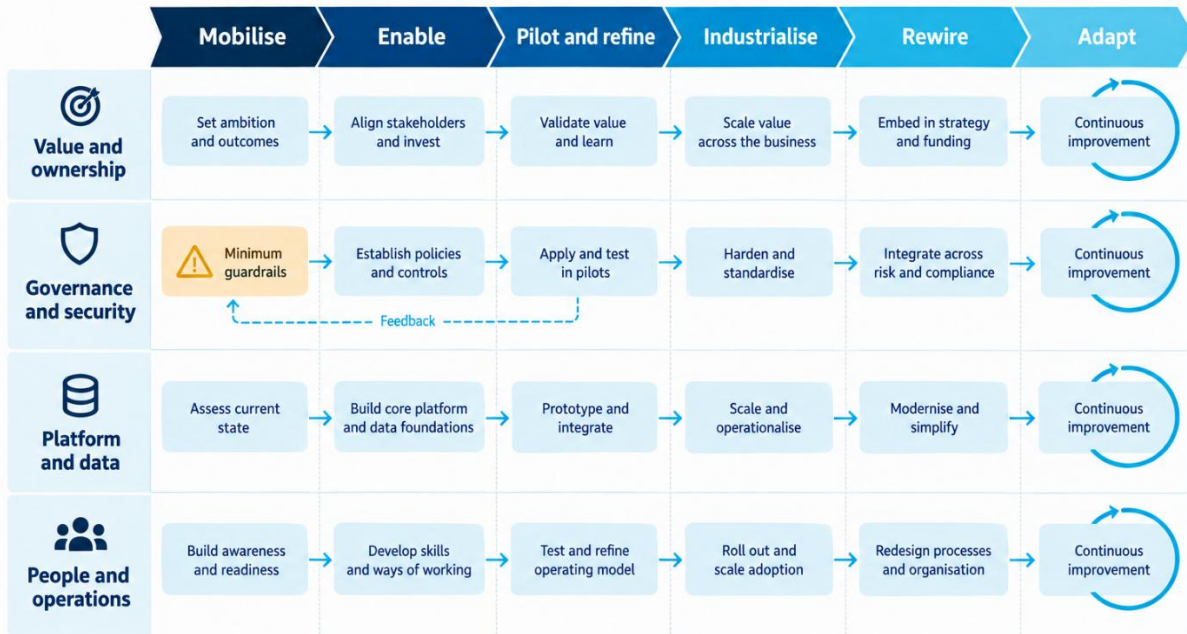
The following is a practical enterprise maturity sequence. The sequence is indicative rather than time-bound; a regulated bank, government institution or critical-infrastructure operator will typically require more assurance than a lower-risk commercial organization.

Before a material use case moves between stages, apply four parallel gates: **value, technical/operational readiness, risk/compliance, and workforce/organizational impact**. The workforce gate asks whether the proposed change alters roles, working conditions, employee monitoring or decision processes sufficiently to trigger employment-law, collective-consultation, works-council or privacy obligations in the relevant jurisdiction.

For Sweden, apply the MBL analysis in the Workforce section; the trigger is the significance of the change — **not the mere fact that AI is involved**. [38] Other jurisdictions require their own assessment.

Stage	Sequence	Primary objective
Mobilize and baseline	Phase 1	Establish mandate, inventory, governance and value thesis.
Controlled enablement	Phase 2	Give employees safe AI access and establish foundational controls.
Governed pilots and standards co-development	Phase 3	Run governed end-to-end pilots while co-developing and refining the minimum standards, controls and reusable delivery patterns required for production.
Industrialise and build federated capability	Phase 4	Standardize reusable delivery patterns and build federated capability under supervised delegation.
Workflow and workforce rewiring	Phase 5	Redesign processes, roles and operating models around AI.
Adaptive AI enterprise	ongoing	AI becomes a normal, continuously managed enterprise capability.

Maturity is coordinated change.



Mobilize and baseline

The first task is not building agents. It is establishing facts.

Create the executive mandate and AI Council; appoint the enterprise AI leader; establish the initial AI Office; inventory existing models, employee AI tools, embedded SaaS AI, shadow AI and automation; map current regulations and national competent authorities; identify employment/worker-consultation triggers; identify major business value pools; assess current data/platform capability; and issue an interim acceptable-use policy.

The baseline should include existing traditional ML because organizations frequently make the mistake of governing GenAI while leaving older AI models in a separate governance universe.

Create the enterprise AI register immediately. Its detailed contents are set out in Section 5.2, Governance and accountability.

Asset inventory is a recurring theme across current AI governance and security guidance; Google's SAIF, for example, explicitly asks organizations whether they maintain complete inventories of models, datasets and associated ML artifacts. [33]

Controlled enablement

Employees should receive an approved enterprise GenAI environment before the organization spends months attempting to eliminate shadow use solely through policy. This is not a theoretical problem: Gartner reported in 2026 that **88% of employees with enterprise AI access also used personal AI tools for business tasks**, reinforcing the case for making governed enterprise AI easier to use than unsanctioned alternatives. [64]

Implement:

SSO, RBAC, approved model access, data-loss controls, logging, contractual privacy protections, usage guidelines, data classification rules, basic cost monitoring, user training, AI support and an intake process for additional tools.

At the same time, establish the formal AI lifecycle, vendor intake processes, workforce-impact assessment and jurisdictional consultation route. Where a proposed deployment materially changes work or introduces employee monitoring, consultation and privacy requirements should be resolved **before** the relevant business decision or deployment gate, not retrospectively.

This is where the requirement to make AI **"easy to consume, safe by default and financially accountable"** is particularly instructive. This calls for a central service model spanning external AI products and internally built services, including assessment, data classification, procurement, identity integration, pilot, rollout, support, documentation and service handover.

Governed pilots and standards co-development

Select a small portfolio containing different archetypes, for example:

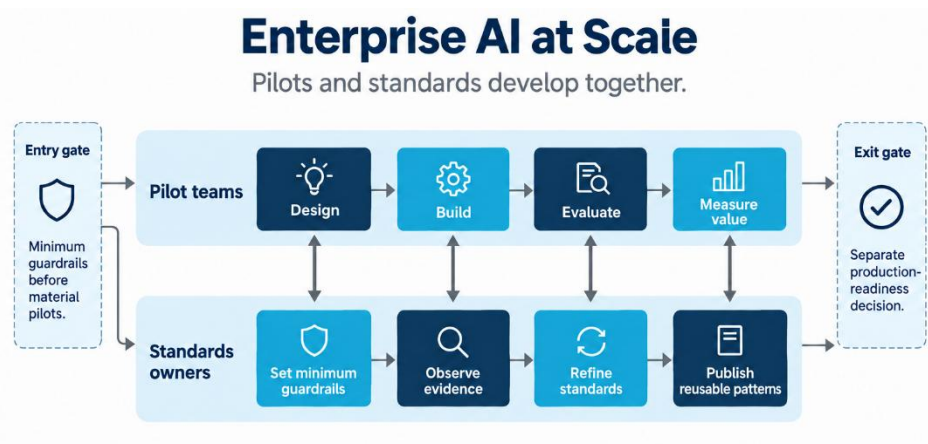
- broad employee productivity;
- one high-volume knowledge workflow;
- one contained customer/service workflow pilot where external exposure is suitably bounded;
- one predictive ML application;
- one bounded agentic workflow.

Each governed pilot should exercise the full intended lifecycle:

discovery → business case → risk classification → workforce/privacy applicability assessment → architecture → controls → build → evaluation → security and systemic testing → consultation/approval where required → business pilot → production-readiness assessment → standards/control refinement → benefits measurement → go/no-go decision.

The purpose is to co-develop and refine the production system around AI from real delivery evidence, without treating the pilot itself as authority to go live.

The starting point should therefore be a minimum standard set before the first material pilot: acceptable use, identity and data boundaries, security requirements, risk/legal/privacy screening, evaluation expectations, supplier controls, escalation and evidence. Pilot teams and the standards owners should then work together continuously so lessons are incorporated into those standards in real time.



Industrialise and build federated capability

During the pilot cycle, productize the capabilities that teams repeatedly need so standards and reusable services evolve with the pilots rather than after them:

- standard architectures;
- model gateway;
- RAG service;
- evaluation framework;
- agent runtime;
- tool registry with ownership, version/change controls and metadata inspection;
- approved connectors and agent protocol patterns;
- AI CI/CD;
- model/agent registry;

- monitoring;
- templates;
- standard contract clauses;
- risk assessment workflow;
- service-onboarding workflow.

Business units may receive delegated authority to deliver within predefined guardrails once the relevant domain capability has demonstrated readiness.

Delegation should be earned progressively rather than handed over in one step. A practical pattern is domain readiness accreditation: demonstrate named accountable roles, trained practitioners, use of approved architecture and controls, budget ownership, evidence quality, incident/escalation routes and successful co-delivery before receiving wider delegated build authority. Different domains can progress at different speeds.



The central team moves from **doing the work** toward **supporting the ecosystem through which work is done**.

Workflow and workforce rewiring

This is where the largest strategic value is likely to emerge.

Instead of asking:

"Where can we insert Copilot?"

ask:

"If this process were designed today around human judgment, deterministic automation, predictive AI and agents, what would it look like?"

Current research increasingly points toward this distinction. Organizations often obtain only incremental gains when AI augments the old workflow; end-to-end redesign, altered decision rights and role changes are required for more transformative outcomes. [21, 22]

A future process might combine:

machine ingestion → deterministic validation → ML classification → agent research → GenAI drafting → policy engine → human exception decision → agent execution → automated reconciliation → continuous monitoring.

That is materially different from giving everyone a chatbot.

Adaptive AI enterprise

At maturity:

AI is included in strategy and annual planning; every material AI asset is known; new AI services have a predictable onboarding route; domain teams can develop safely without waiting for central experts; data is exposed through

governed reusable products; humans and agents have explicit decision rights; model and agent performance is monitored; costs are attributable; benefits are auditable; obsolete assets are retired; incidents improve the control framework; and employees continuously develop AI capabilities.

The mature enterprise is therefore **AI-pervasive but not AI-maximalist**.

"AI first" should mean *consider AI as a normal design option, not use AI whether or not it is appropriate*.

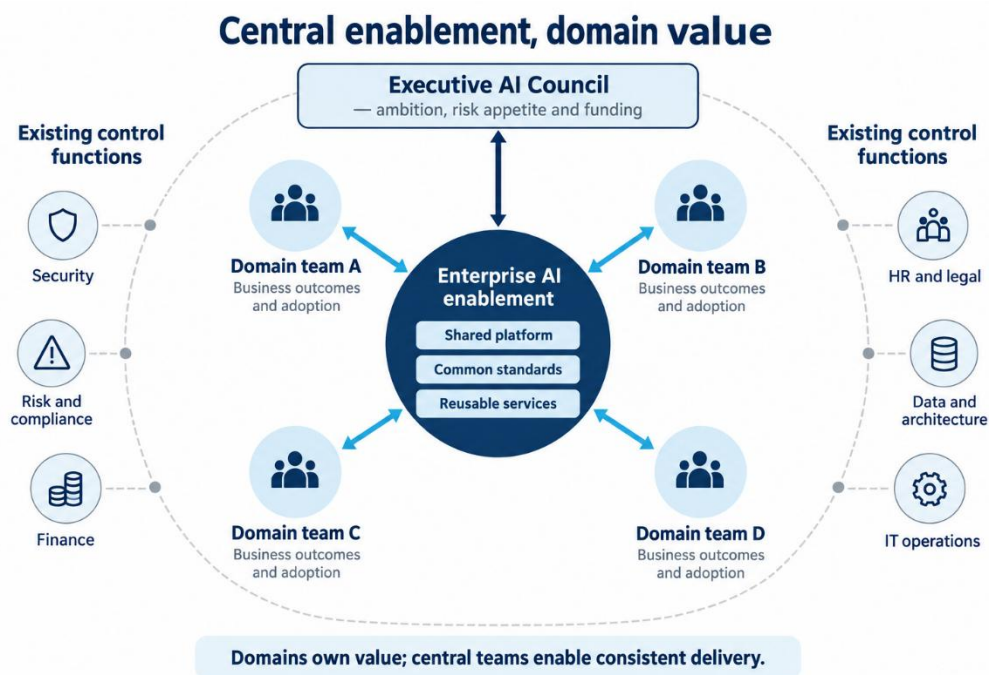
PART III — THE CAPABILITY MODEL: WHAT MUST MATURE TOGETHER

Target-state operating model, roles and technology foundation

The strongest enterprise architecture is a **central control plane plus decentralized value creation**.

The central AI function should *not* become the organization that builds every AI application. Its purpose is to make good AI solutions easier to build. Domain teams should be able to consume approved models, data, tools, evaluation services, agent capabilities and deployment patterns through paved roads rather than negotiating governance from first principles for every project.

This federated arrangement closely matches the large-enterprise CoE pattern described by IBM and the strategic AI council model advocated in recent enterprise AI research. McKinsey recommends cross-functional strategic oversight of AI, IT and data investment with rigorous tracking against business KPIs, while simultaneously placing lighthouse transformations in the operating business. [3, 4]



3.1 Twelve capability domains

For navigation, this paper uses the same twelve domains in the high-level capability model, the maturity matrix and the detailed guidance. The domains are not twelve new departments. They are accountability areas that should normally map onto — and extend — the organisation's existing operating model.

Capability domain	What it covers	Likely existing enterprise home
Strategy / value	AI ambition, value pools, portfolio, business cases and benefits	Strategy, transformation, business leadership, Finance
Governance	Decision rights, risk appetite, AIMS, policy, exceptions and regulatory coordination	Existing governance, risk, compliance, model-risk and executive forums
Security	Identity, least privilege, action boundaries, threat modelling, adversarial resilience and incident response	CISO, InfoSec, IAM, DevSecOps, SOC
Platform	Approved models, gateways, shared AI services, tooling, observability and reusable architecture	Technology platform, cloud, enterprise architecture

Capability domain	What it covers	Likely existing enterprise home
Data	Structured/unstructured data, RAG, knowledge, lineage, memory and lifecycle controls	CDO/data governance, records, data engineering
Delivery	AI SDLC, product delivery, engineering patterns, MLOps/LLMOps/AgentOps and release management	Engineering, DevOps, product/delivery
Operations	ITSM, service ownership, support, continuity, change, incident, problem and retirement	Service management and operations
People	Role capacity, specialist capability, training, champions and communities of practice	HR/L&D, AI Office/CoE, domain leadership
Economics	Funding, FinOps, cost attribution, budgets, unit economics and benefit conversion	Finance, FinOps, portfolio management
Workforce / legal	Role redesign, consultation, employee monitoring, privacy and legal obligations	HR/employee relations, Legal, Privacy, Compliance
Systemic assurance	Risk tiers, autonomy, evaluations, red teaming, human oversight and system-level evidence	QA/testing, model risk, second line, Internal Audit
Vendor / third-party	Supplier intake, contracts, AI supply chain, version change, concentration and exit	Procurement, vendor management, third-party risk

Security and systemic assurance are cross-cutting even though they are shown as separate domains for accountability and maturity measurement. Their placement in a table does not imply that they can be sequenced after build activity.



PART IV — MATURITY ACROSS THE CAPABILITY DOMAINS

The maturity model turns the capability map into an evidence-based progression. It is intended to help an organisation identify where it is today, where it wants each domain to reach, and what observable evidence should exist before it claims to have moved forward.

Assess each capability on evidence

● Circle = current ◆ Diamond = target

Capability	Experimental	Controlled	Standardized	Federated	AI-pervasive
Strategy/value					
Governance					
Platform					
Data					
Delivery					
Operations					
People					
Economics					
Workforce/legal					
Systemic assurance					
Security					
Vendor/third-party					

Assessment template — no organisation scores supplied.

A practical maturity model

The target state can be assessed across **twelve dimensions**:

Dimension	Experimental	Controlled	Standardized	Federated	AI-pervasive
Strategy/value	Ad hoc ideas	Selected pilots	Formal portfolio	Domain portfolios linked to strategy	AI integral to corporate planning
Governance	Policies absent	Interim controls	Standard lifecycle	Automated/risk-tiered governance	Continuous evidence and assurance
Platform	Individual tools	Approved tools	Common platform	Self-service platform	Multi-model/agent control plane
Data	Case-specific extracts	Controlled access	Reusable pipelines	Governed data/knowledge products	AI-ready data by design
Delivery	PoCs	Lighthouse projects	Repeatable AI SDLC	Domain product teams	Continuous AI product delivery
Operations	No ownership	Manual support	MLOps/LLMOps	AgentOps + ITSM	Fleet-level automated operations
People	Enthusiasts	Basic literacy	Role training	Champions/domain skills	Human-AI role design embedded
Economics	Innovation spend	Pilot business	Benefits tracking	Portfolio optimization + runtime budgets	Cost/value telemetry and

Dimension	Experimental	Controlled cases	Standardized	Federated	AI-pervasive enforcement by outcome
Workforce/legal	Unassessed impacts	Ad hoc HR/legal involvement	Standard impact/consultation gate	Domain workforce redesign with formal evidence	Workforce governance integrated into planning and operating model
Systemic assurance	Component demos	Basic model/app tests	Standard evaluation + security testing	Multi-agent/tool-chain fault and adversarial testing	Continuous system-level assurance, replay and automated containment
Security	AI-specific risks unmanaged	Basic access and review	ISMS-integrated AI security standard, threat model and red-team practice	Policy-enforced agent/tool controls and fleet telemetry	Continuous threat intelligence, automated containment and measurable control effectiveness
Vendor/third-party	Ad hoc tool purchase	Approved supplier list	Standard due diligence, contracts, version/change controls and exit requirements	Portfolio-level concentration, portability and supply-chain governance	Continuous dependency intelligence, tested switching and resilient multi-provider options where justified

The goal is not to maximize the number of models or agents. The goal is an organization where **AI is routinely considered, rapidly deployable when appropriate, provably controlled, economically transparent and capable of changing how work is performed.**

Maturity should be evidenced, not only self-declared. Suggested **stage-transition evidence** is:

Transition	Minimum observable evidence	Sampleable evidence pack
Experimental → Controlled	Enterprise AI inventory, named owners, acceptable-use rules, approved tool access, baseline literacy, initial risk/intake process.	Inventory export with owners; approved-use/policy record; training completion evidence; sample IAM/tool-access logs; sample risk/intake decisions.
Controlled → Standardized	Common lifecycle/risk tiers, reusable platform patterns, evaluation/security standards, vendor process, workforce-impact gate, production OAT/ITSM and AIMS evidence cadence.	A sample of end-to-end use cases showing business case, risk/impact decision, architecture/security/privacy/workforce decisions, versioned evaluation results, supplier record, CMDB/OAT evidence and management-review actions.
Standardized → Federated	Documented central/domain decision rights, delegated delivery within guardrails, automated policy/evidence capture, domain value ownership,	Delegation/RACI records; policy-engine/authorization logs; repeatable release evidence; domain benefit realization; cost showback/chargeback; exceptions and corrective actions sampled across domains.

Transition	Minimum observable evidence	Sampleable evidence pack
	standardized cost attribution and repeatable production releases.	
Federated → AI-pervasive	Continuous fleet-level assurance, automated containment/revocation, outcome-level economics, management review/internal audit, resilient multi-provider operations, measurable workflow/workforce redesign and retirement of low-value AI.	Fleet assurance dashboard; internal-audit sample; incident/nonconformity and corrective-action trail; failover/continuity exercise; realized outcome economics; workforce/process redesign evidence; retirement/decommission records.

No organization should advance a maturity label merely because usage volume increased. The evidence must show stronger **outcomes, controls, repeatability and operating capability**.

The executive scorecard

A mature AI program should report a small set of outcome metrics rather than hundreds of AI activity measures.

Value

- annualized realized benefit versus approved business cases;
- revenue/margin contribution;
- productive capacity released and actually redeployed;
- quality/error improvements;
- value per £/€/SEK of AI run cost.

Portfolio

- number/value of ideas, experiments, production products and retired use cases;
- conversion rate from pilot to scaled production;
- time from approved opportunity to production;
- percentage of portfolio benefits concentrated in strategic value pools.
- percentage of scaled use cases with completed six- and twelve-month post-scale benefit audits.

Adoption

- eligible versus active users;
- sustained weekly/monthly usage;
- workflow penetration rather than login rate;
- user trust/satisfaction;
- percentage of relevant population trained.

Risk

- percentage of AI estate inventoried and owned;
- percentage current on risk review/evaluation;
- open high-severity control deficiencies;
- AI/security incidents and near misses;
- agent permissions outside standard policy;
- unapproved tool/server versions or protocol configurations;
- workforce/privacy consultation gates overdue;
- number of overdue reassessments.
- open exceptions/waivers and their age;
- percentage of production agents whose consequential actions traverse a policy-enforced API or equivalent deterministic boundary.

Performance

- task-specific evaluation scores;
- groundedness/accuracy where measurable;
- human override/exception rate;
- latency/reliability;
- model drift;
- agent task-success and unsafe-action rate.

Economics

- model/inference spend;
- cost per transaction/task/**successful outcome**;
- token/context consumption and tool/API spend by workflow;
- execution-budget/circuit-breaker breaches;
- unused licenses;
- cost by domain/use case/provider;
- forecast versus actual run cost.

Operational health

- incidents;
- mean time to detect/contain/recover;
- service availability;
- model/agent version currency;
- time required to disable an agent, tool/server or revoke access;
- multi-agent workflow failure containment/recovery;
- RAG/index deletion and access-change propagation health.
- percentage of AI-generated production-code changes that passed protected-branch, CI and independent-review controls.

The most important metric is ultimately:

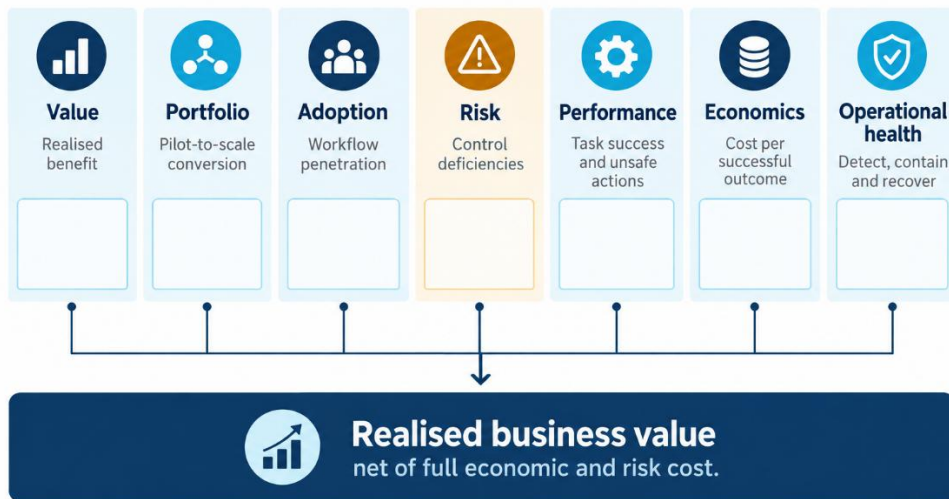
Realized business value from production AI, net of its full economic and risk cost.

The overarching implementation principle for a large institution can therefore be expressed in one sentence:

Centralize the capabilities that must be consistent—governance, security, platform, standards, economics and reusable assets—while federating the work that must remain close to the business: process ownership, product design, domain knowledge, change and accountability for value.

That operating model directly addresses the strongest themes running across the internal job-specification corpus while also aligning with current standards and evidence on enterprise AI, GenAI and agentic-AI deployment. It provides the route from today's fragmented AI experiments to a mature organization in which traditional ML, generative AI and increasingly autonomous agents operate as normal, measurable and governed components of the enterprise rather than as isolated technology initiatives. [1, 2, 3, 4, 5]

Measure outcomes, not AI activity.



PART V — DETAILED GUIDANCE BY CAPABILITY DOMAIN

Part V is the lowest level of detail in this umbrella paper. It groups the existing guidance by the capability domains introduced in Part III. Future domain papers can deepen any one area without changing the overall framework.

5.1 Strategy and value

Business cases, use-case selection and portfolio management

This is arguably the most important part of enterprise AI management because the principal failure mode is no longer inability to build a demo. It is **building things that work technically but do not generate material enterprise value**.

Academic evidence confirms that GenAI can generate meaningful productivity gains, but also demonstrates why organizations should not substitute published productivity percentages for their own business cases.

In a study of **5,172 customer-support agents** published in the *Quarterly Journal of Economics* in 2025, access to a GenAI assistant increased issues resolved per hour by about **15% on average**, with much larger gains among less experienced workers. A separate experiment involving 758 consultants found that, for tasks inside the model's capability frontier, users completed work more quickly and at higher assessed quality; importantly, the study also found a "jagged frontier" where AI performance varied by task. A 2025 field experiment involving 7,137 knowledge workers across 66 firms found that users of integrated GenAI tools saved roughly two hours of email time per week in the latter half of the trial but did not show broader changes in work composition merely from receiving individual-level AI tools. [10, 69, 11, 12]

The counter-evidence matters just as much. METR found that experienced open-source developers in its 2025 randomized study were **slower** with then-current AI tools despite expecting to be faster, and its February 2026 follow-up still did not establish the large speedups often assumed in business cases. Humlum and Vestergaard found precise near-zero average effects on earnings and hours across exposed occupations while still observing modest self-reported time savings, and BetterUp/Stanford research on "workslop" illustrates how low-quality AI output can shift effort to recipients rather than remove it. These studies do not prove that enterprise AI lacks value; they show why **local measurement, workflow redesign and quality/rework accounting are mandatory**. [70, 71, 72, 73]

The practical conclusion is:

The unit of investment should normally be the business workflow or product outcome, not the AI tool.

Start with value pools, not an AI idea competition

The organization should initially map the largest economic and strategic value pools:

Cost and productivity: labor-intensive workflows, administration, documentation, service, technology delivery.

Revenue and growth: sales effectiveness, personalization, product innovation, conversion and speed to market.

Quality: defects, rework, inconsistency, processing accuracy, decision quality.

Risk: fraud, financial crime, cyber risk, compliance effort, loss avoidance.

Customer/employee experience: response time, service quality, knowledge access and friction.

Capital and operational performance: inventory, forecasting, scheduling, maintenance and optimization.

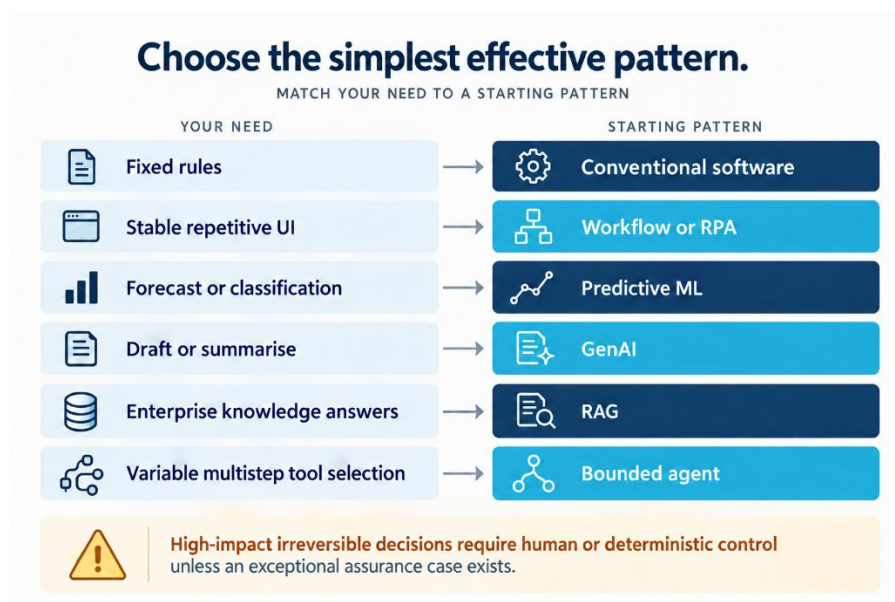
Only after identifying these should teams ask how AI helps.

Select the right form of automation

A mature organization should explicitly challenge requests for "an agent":

Problem	Preferred starting pattern
Fixed logic and repeatable rules	Conventional software / rules engine
Stable repetitive UI workflow	RPA/workflow automation
Forecast or classification	Predictive ML
Optimization under constraints	Mathematical optimization / ML
Draft, summarize, translate, extract	GenAI
Answer from enterprise knowledge	RAG-enabled GenAI
Reason within a deterministic process	GenAI embedded in workflow
Variable multistep process requiring tool selection	Bounded agent
High-impact irreversible decision	AI recommendation plus human/deterministic control, unless an exceptional assurance case exists

Agentic AI should therefore be **earned by the use case**, not adopted as the architectural default.



Apply a risk gate before economic ranking

There should be two separate decisions.

First determine whether the use case is permissible and what controls it requires:

- prohibited or restricted use, including enterprise-relevant Article 5 checks such as manipulative/exploitative techniques, social scoring, workplace/education emotion inference (subject to the medical/safety exception), prohibited biometric categorisation and other applicable prohibited practices;
- AI Act classification;
- personal/sensitive data implications;
- legal or fundamental-rights impact;
- security classification;
- consequential decision-making;
- external/customer interaction;
- financial authority;
- irreversible actions;
- agent access privileges;

- sector-specific requirements.

Only then rank viable projects economically.

This matters particularly in Europe because the EU AI Act is already substantially in operation. The Regulation entered general application on **2 August 2026**, but its obligations remain phased: **all original Article 5 prohibitions** and the Article 4 AI-literacy duty have applied since 2 February 2025, with penalty provisions applying from 2 August 2025; governance and GPAI obligations since August 2025; Article 50 transparency from August 2026; the new Article 5 prohibitions introduced by the **Digital Omnibus on AI (Regulation (EU) 2026/1744)** apply from **2 December 2026**; relevant Annex III high-risk obligations apply from **2 December 2027**; and high-risk systems embedded in regulated Annex I products apply from **2 August 2028**. [6, 7, 8, 35, 36]

Use an explicit prioritization score

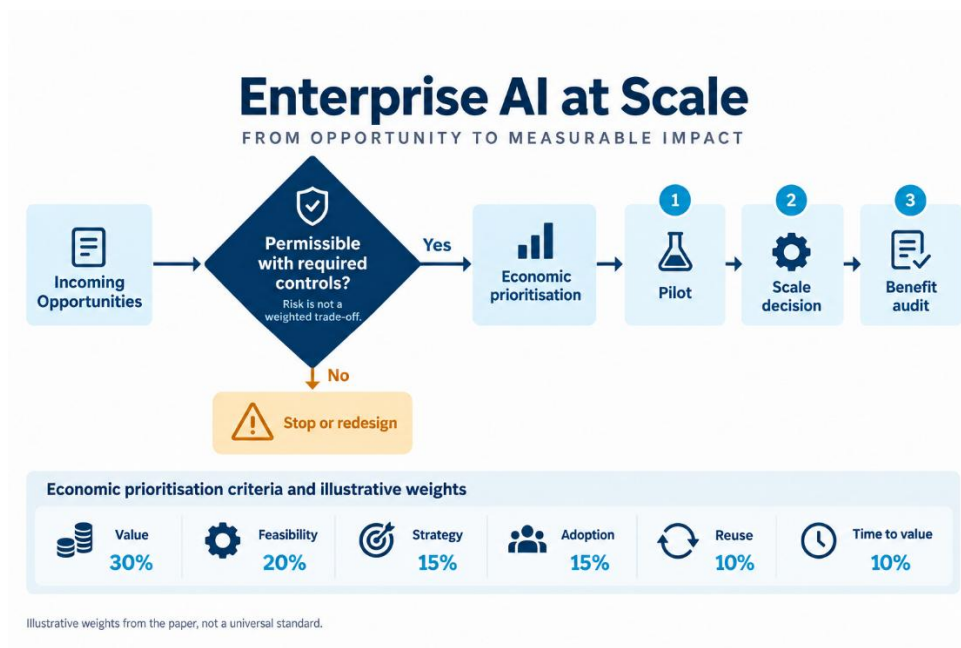
A useful starting portfolio score—not a universal standard—is:

Dimension	Suggested weight
Expected business value	30%
Technical and data feasibility	20%
Strategic importance	15%
Workflow/adoption readiness	15%
Reusability and enterprise scalability	10%
Time to measurable value	10%

Risk is deliberately **not** just another weighted variable. Certain risks require controls or make the project impermissible regardless of economic attractiveness.

Each candidate should have a one-page opportunity card containing:

business owner; problem; workflow; baseline; target outcome; users/customers; proposed AI pattern; data required; action authority; regulatory/risk classification; dependencies; implementation estimate; expected benefit; full cost; capacity/value conversion mechanism and named owner where relevant; measurement design; exit/rollback strategy.



Build the business case from an operational baseline

For productivity cases, a useful starting equation is:

Annual capacity value = annual transaction volume x minutes saved per transaction x loaded labor cost per minute x realizability factor.

The **realizability factor** is critical. Saving an employee five minutes does not automatically reduce the company's cost by five minutes of salary. The business case must state how released capacity will actually be converted into value: more throughput, lower outsourcing, avoided hiring, faster turnaround, redeployment, revenue activity or staffing reduction.

For productivity cases, name the **capacity-conversion mechanism owner** at approval — normally the operating manager who controls staffing, throughput, outsourcing, service levels or redeployment. A scale decision should use **measured conversion evidence** where the pilot has run long enough to observe it; projected time savings alone are not sufficient proof of realized benefit.

Total benefit may also contain:

Benefit = capacity + rework avoided + loss/risk avoided + incremental margin + working-capital effect + quality/experience value.

Against that, use **full lifecycle TCO**, not model API cost:

- licenses/subscriptions;
- tokens/inference/compute;
- data acquisition and remediation;
- integration;
- platform services;
- cybersecurity;
- governance/compliance;
- evaluation/red teaming;
- implementation;
- training and change;
- human review;
- support and operations;
- model/agent monitoring;
- vendor management;
- technical debt;
- exit/migration cost.

For agents, add the **cost of failure and control**. An apparent labor-saving case may cease to be attractive if every transaction requires expensive human checking. Conversely, workflow redesign can make a case substantially stronger by moving human judgment to exceptions instead of requiring review of every step.

Business cases should use ranges rather than false precision: conservative, base and upside scenarios, preferably with explicit assumptions.

The model bill is only part of the cost.



Model agentic economics at the workflow level

The capacity equation above is appropriate for many copilot and task-augmentation cases. It is not sufficient for an agentic workflow whose value proposition is to remove handoffs, restructure work or automate an end-to-end process. Gartner's 2025 warning that more than 40% of agentic AI projects could be cancelled by the end of 2027 specifically cited escalating cost, unclear business value and inadequate risk controls, and recommended using agents where they deliver clear enterprise productivity rather than merely individual task augmentation. [48]

Agentic business cases should therefore measure the **whole operating process**. Useful variables include: end-to-end cycle time; handoffs removed; straight-through-processing rate; exception and escalation rate; human review minutes per completed case; rework; failure/recovery cost; tool/API/model consumption; risk/control effort; service capacity; and cost per successful business outcome.

A useful expression is:

Net agentic value = process value created - run cost - human exception/control cost - expected failure loss - change and operating cost.

This forces the organization to compare an agentic design with simpler alternatives such as conventional workflow, rules, RPA or an AI-assisted process.

Pilot to learn, not merely to prove that the technology works

A pilot should answer five questions:

- Does it work technically?
- Is it safe enough?
- Do users change behavior?
- Does it improve the business metric?
- Are the economics attractive at production scale?

Before deployment, capture the baseline. During the pilot, compare AI-enabled performance with the baseline or a suitable comparison group. Measure quality as well as speed.

A scaling decision should require predefined thresholds and a credible measurement design. At minimum, material pilots should define: **a comparison design** (control, matched cohort, phased rollout or defensible baseline); **minimum observation duration**; **quality/rework and downstream burden measures** as well as speed; **novelty-decay and volume-rebound checks**; and the business mechanism by which measured capacity becomes realized value.

After scale, perform a **benefit audit at approximately six and twelve months** for material initiatives, comparing approved assumptions with realized throughput, cost, quality, risk and adoption. A production use case that misses its benefit target should be **reworked or stopped**, even if technically impressive.

Run one portfolio, not separate "AI innovation" and "real IT"

The mature organization should maintain an enterprise AI portfolio covering:

- employee copilots;
- domain-specific GenAI products;
- traditional ML models;
- intelligent automation;
- embedded customer/product AI;
- agentic workflows;
- experimental/emerging AI.

The portfolio should distinguish **exploration budget** from **scale investment**, so experimentation remains possible without allowing perpetual proofs of concept.

McKinsey's current recommendation to concentrate on a limited number of high-impact lighthouse transformations while simultaneously building governance, data, technology and workforce foundations is a sound starting principle. [4]

5.2 Governance and accountability

The enterprise AI register should eventually contain:

- owner;
- purpose;
- system/type;
- model/provider and approved version(s);
- users/affected persons;
- data classes;
- integrations;
- risk and legal classification;
- agent identity, delegated authority and permissions where applicable;
- production status;
- controls and policy-set/version;
- evaluations and evaluation-dataset/version;
- incidents;
- linked risk/impact/DPIA records and transparency/disclosure method where applicable;
- review date;
- cost;
- benefit metrics;
- retirement status.

Governance, policy, security and service introduction

The governance system should be based on a hierarchy of **principles → policies → standards → controls → procedures → evidence**, rather than attempting to place everything in one large "Responsible AI Policy."

The hierarchy should also be **tailorable**. Enterprise principles, accountability, inventory and prohibited boundaries are mandatory; standards define risk-tiered minimums; procedures may be implemented locally where that does not weaken the required control outcome. An artefact is valuable because it supports a decision, control or audit trail — not because the organization can count another document.

ISO/IEC 42001 is particularly useful as the umbrella because it establishes an auditable AI management-system concept based on continual improvement. NIST AI RMF can then provide a practical risk-management structure, while the NIST Generative AI Profile, OWASP guidance and sector-specific standards deepen particular technical areas. [1, 13, 2]

From intent to operating evidence



The organizational model

At the top should be an **Executive AI Council**, chaired by a genuinely accountable business executive rather than functioning as an IT committee. Depending on the organization this could be the CEO, COO, CIO/CDO or a Chief AI Officer. Its permanent representation should normally include technology/data, business operations, finance, HR, legal/privacy, risk/compliance and cybersecurity.

Its job is to decide:

- enterprise ambition and risk appetite;
- priority value pools and transformation domains;
- investment allocation;
- exceptional/high-risk deployments;
- major technology and sourcing choices;
- workforce implications;
- unresolved cross-organizational conflicts.

Below it sits an **Enterprise AI Office / AI Center of Excellence / AI Enablement function**.

Its mandate should include both value and control:

Central capability	Core accountability
AI strategy and portfolio	Target state, roadmap, portfolio prioritization, investment cases, benefits realization and executive reporting.
AI governance and Responsible AI	Policy, AI inventory, risk classification, control framework, lifecycle gates, EU AI Act coordination, standards and assurance evidence.
AI platform	Model access, gateways, developer tooling, agent infrastructure, evaluation, deployment and observability.
Data and knowledge enablement	Data access patterns, RAG/knowledge services, metadata, lineage, unstructured-data governance and reusable data products.
AI security	Threat modelling, IAM, agent identities, secure tool access, DLP, red teaming, supply-chain assurance and incident response.
Engineering excellence	Reference architectures, MLOps/LLMOps/AgentOps, reusable components, testing and production engineering.
AI enablement and change	Literacy, specialist training, champions, communities of practice, adoption, workflow redesign and communication.
AI service management	Intake, onboarding, service catalogue, configuration records, service ownership, support models, incident/change management and decommissioning.
AI economics / FinOps	Model and token spend, licenses, compute, unit economics, showback/chargeback and

Central capability	Core accountability
	vendor economics.

Business units then operate **domain AI teams or spokes**. Each important business domain should have an accountable AI/product lead rather than simply submitting requests to IT. These teams own process knowledge, business requirements, adoption and measurable outcomes while consuming centrally provided capabilities.

Funding should follow accountability. Shared foundations such as the enterprise AI platform, common controls and central enablement can be funded centrally, while business implementations should normally be owned within the relevant domain budget and pass through existing domain investment governance. The Executive AI Council should align ambition with an explicit funding envelope rather than create unfunded strategy.

A mature domain team will typically combine a **business/process owner, AI product owner/s, business analyst/s, data/AI engineers, solution architect and change/adoption lead**, drawing centrally on security, privacy, legal, governance and platform specialists.

The organization should distinguish three accountabilities that are frequently blurred:

The business/process owner owns the business outcome.

The AI/product owner owns the AI-enabled product or service.

The central AI function owns the enabling system and enterprise guardrails.

That distinction prevents the AI CoE from becoming accountable for benefits that only the operating business can realize.

Decision rights for a federated AI operating model

A federated model only works when decision rights are explicit. The following is a minimum decision-rights pattern; enterprises should map the labels to their existing governance structure rather than create duplicate committees.

Decision	Accountable decision owner	Mandatory challenge / contribution	Required evidence
Enterprise AI ambition and risk appetite	Governing body / executive management	Executive AI Council; second line	Strategy, risk appetite, approved funding envelope, material exposure and value portfolio.
Use-case approval and production tier	Domain/process owner within delegated authority; Executive AI Council for exceptional/critical cases	AI Office; security/privacy/legal/model-risk as applicable	Opportunity card, classification, impact/risk assessment and release evidence.
Grant or increase consequential action authority	Named business/process owner within approved risk limits	AI Office/platform security; second line for material/critical authority	Autonomy decision, least-privilege design, negative authorization tests, rollback/kill evidence.
Accept material residual risk	Formally delegated risk owner; executive authority where outside tolerance	Independent second-line challenge	Residual-risk statement, compensating controls, expiry/review condition.
Approve an exception/waiver	Authority defined by risk tier; no self-	Control owner and second line	Scope, rationale, compensating

Decision	Accountable decision owner	Mandatory challenge / contribution	Required evidence
	approval by the control owner		controls, expiry, owner and register entry.
Stop, scale or retire a use case	Domain/process owner for value; Executive AI Council for strategic/material portfolio decisions	AI Portfolio & Value Lead; service owner; risk input where material	Benefit evidence, control/incident record, run cost and exit implications.
Approve material model/provider/tool/scope change	Service/product owner under change authority	Architecture/security/privacy/risk specialists triggered by impact	Change assessment, regression evidence and updated configuration record.
Declare and coordinate a material AI incident	Existing enterprise incident authority / service owner	AI Office, security, privacy/legal, regulatory reporting owners	Incident classification, containment, notification clock and evidence trail.

Internal Audit does not approve first- or second-line decisions. It independently assesses whether the AIMS, decision rights and material controls are designed and operating effectively.



The governing body should receive a proportionate view of **realized value, material AI risks, significant incidents, risk-appetite breaches, overdue exceptions, regulatory readiness and AIMS effectiveness**. The organization should state an AI risk appetite in operational terms — for example, which decisions may never be solely automated, which authority levels agents may hold, which data classes may be exposed to which providers, and which residual-risk decisions require executive approval.

Scale the operating model proportionately

The target state is deliberately comprehensive, but **comprehensive does not mean every control, committee or artefact must be implemented at maximum intensity for every organization or every use case**. ISO/IEC 42001 and NIST AI RMF both begin from organizational context, scope, risk and intended use. [1, 2] A mid-sized commercial company, a regulated bank and a critical-infrastructure operator can therefore share the same control objectives while implementing them with different depth, independence and evidence burden.

A practical rule is to separate **always-on control objectives** from **risk-tiered assurance intensity**:

Always-on objective	What may scale with risk, size and sector
Every material AI service/use case has a named owner and inventory record.	Number and seniority of governance forums; independent review depth.
Approved data, identity and access boundaries are explicit.	Formal architecture/security/privacy documentation and specialist sign-off.
The use case is risk-classified before production.	Evaluation breadth, red teaming, parallel run and independent assurance.
Consequential authority and spend are bounded deterministically, and consequential actions traverse a policy-enforced boundary wherever technically feasible.	Transaction ceilings, approval layers, monitoring coverage, isolated execution surfaces and 24/7 operations.
Costs and intended benefits are attributable.	Chargeback sophistication, portfolio analytics and FinOps automation.
Incident, revocation, rollback and retirement paths exist.	Resilience testing frequency, disaster-recovery depth and dedicated support model.
Workforce/legal/privacy applicability is assessed.	Consultation, impact-assessment and evidence requirements according to jurisdiction and use.

A **minimum-viable enterprise AI operating model** should contain at least:

1. **Accountability:** executive sponsor, named business/process owner for each material use case, and a clear central owner for platform/guardrails.
2. **Inventory and policy:** enterprise AI register, acceptable-use rules, risk/autonomy tiers and a defined exception route.
3. **Approved technical path:** managed model/tool access with identity, logging, data controls, deterministic policy boundaries and cost metering.
4. **Lifecycle:** intake → value case → risk/data/security/privacy/workforce applicability → evaluation → production/service acceptance → monitoring → retirement.
5. **Operations:** service/support owner, incident route, access revocation, rollback/kill path, minimum telemetry and cost controls.
6. **People:** baseline AI literacy plus role-specific training for builders, approvers, managers and risk/control functions.

Organizations can then add specialist roles and artefacts as exposure grows. The objective is to avoid two failure modes: **under-governed AI because the full model looks too heavy, and governance theatre because every low-risk use case is forced through a bank-grade process.**

Operate governance as an AI management system

ISO/IEC 42001 should function as the **management-system backbone**, not merely as one item in a list of frameworks. ISO defines it as a system for establishing, implementing, maintaining and continually improving an Artificial Intelligence Management System (AIMS), using a Plan-Do-Check-Act approach. [1]

The blueprint should therefore be auditable across the management-system sequence: **organizational context and scope; leadership/accountability; planning and risk treatment; resources/competence/documentation; operational lifecycle controls; monitoring/audit/management review; and corrective action/continual improvement.** The practical implication is that the artefacts in this document — policy, register, impact assessments, evaluations, incident evidence, metrics and management decisions — are not disconnected paperwork. Together they form the operating evidence of the AIMS.

Where an organization seeks formal ISO/IEC 42001 certification, exact clause and Annex A mappings should be completed against the licensed standard and the organization's selected scope. The blueprint should not pretend that a secondary web summary substitutes for the normative text.

A management system becomes auditable when evidence is repeatable, owned and reviewed. A practical **AIMS evidence pack** should include at least:

Evidence class	Typical evidence
Scope and context	AIMS scope, interested parties, applicable legal/contractual requirements and material AI estate.
Leadership and objectives	AI policy, accountable roles, approved objectives/KPIs, decision rights and resourcing decisions.
Risk and impact	AI risk assessments/treatment decisions, exceptions, system impact assessments and reassessment triggers. ISO/IEC 23894 provides AI risk-management guidance and ISO/IEC 42005 provides a structured AI system impact-assessment framework. [56, 57]
Competence and awareness	Role requirements, AI-literacy curricula, completion/competence evidence and targeted remediation.
Operational control	Lifecycle gates, design/security records, evaluations, change records, service acceptance and supplier evidence.
Performance evaluation	Control/KPI monitoring, internal audit programme/results, trend analysis and assurance findings.
Management review	Periodic management-review inputs, decisions, actions, owners and follow-up.
Improvement	Incidents/nonconformities, root-cause analysis, corrective action, lessons learned and verified closure.

This is an implementation pattern, not a substitute for the licensed ISO/IEC 42001 text. It also creates a natural crosswalk to NIST AI RMF **Govern → Map → Measure → Manage** rather than running two parallel governance systems. [1, 13]

A large enterprise should also make the **Three Lines** assurance relationship explicit rather than leaving it implicit in the org chart. The IIA's current Three Lines position emphasizes governing-body oversight, first- and second-line management/risk roles and independent third-line assurance. [67]

Line / governance role	AI accountability
Governing body / executive oversight	Sets AI ambition and risk appetite; receives evidence on value, material risk, incidents and management-system effectiveness.
First line — business/product/engineering/service owners	Own AI-enabled outcomes, operation, controls, evidence, incidents and benefits within approved guardrails.
Second line — risk/compliance/privacy/security/model-risk functions	Set or interpret control standards, independently challenge classification/assurance, monitor exposure and advise on risk acceptance.
Third line — Internal Audit	Independently assesses the design and effectiveness of the AIMS and material AI controls without owning them.

The mapping should be tailored to the organization's existing governance model; creating a separate AI risk bureaucracy is not the objective.

Board/executive oversight should use the decision-rights model defined earlier rather than becoming a parallel approval layer. At minimum, the governing body or its delegated committee should periodically receive the AI risk-appetite position, material portfolio value, significant incidents, overdue/high-impact exceptions, major regulatory-readiness gaps and evidence on AIMS effectiveness. Where AI controls are material to the organization's broader control framework, they should be included in the normal governance and assurance process rather than reported as technology activity. [67, 86]

The required policy and control library

A serious enterprise program should ultimately maintain at least the following artefacts:

Artefact	Purpose
Enterprise AI Policy	Principles, scope, accountabilities, mandatory governance.

Artefact	Purpose
Acceptable AI Use Standard	What employees may and may not do with enterprise and public AI tools, with named Article 5 screening categories and restricted content/use classes rather than a generic "comply with law" statement.
AI Risk Classification Standard	Enterprise risk tiers and relationship to legal classifications.
AI System Impact Assessment Procedure	Structured assessment of intended/unintended effects on individuals, groups and society, with lifecycle reassessment; align to ISO/IEC 42005 where appropriate. [57]
AI Act FRIA Procedure	Applicability test, assessment, market-surveillance notification and lifecycle update for deployments legally in scope of Article 27; separate from but able to cross-reference DPIA. [35]
AI Lifecycle / AI SDLC Standard	Required controls from discovery through retirement.
AI Use-Case and System Register	Authoritative inventory and ownership.
Data for AI Standard	Data classification, personal/confidential data, training/fine-tuning/synthetic/RAG data, licensing/consent where applicable, content ownership/freshness, derived embeddings/indexes, retention, lineage and erasure propagation. Where Article 4a is relied on for bias detection/correction using special-category data, record strict necessity and apply its cumulative safeguards; it does not create an obligation to process such data. [36, 101, 102]
GenAI Development Standard	Prompting, grounding, output handling, evaluation and model use.
Agentic AI Standard	Identity, delegation, autonomy, tools, authority, memory/state, approvals, inter-agent boundaries and runtime controls, explicitly including coding, infrastructure, CI/CD and UI-acting agents.
AI Security Standard	Threat modelling, secure development, red teaming, model/tool/protocol supply chain, sandboxing and monitoring; cross-reference the enterprise ISMS/secure-development baseline and use AI threat taxonomies such as NIST AI 100-2/MITRE ATLAS where useful. [87, 88, 89, 90, 91]
AI Evaluation Standard	Required quality, safety, robustness, security, bias/fairness and regression tests.
Human Oversight / Automated Decision-Making Standard	What requires approval, review, meaningful human intervention, information, contest/appeal, override and escalation; include GDPR Article 22 applicability and AI Act Articles 14(4)(b), 26(2) and 86 where relevant. [59, 35, 80, 81, 82]
Transparency Standard	Human-AI interaction disclosure, Article 50(3) emotion/biometric deployer information, synthetic-content marking/labelling and user communications; use the Commission Code/Guidelines as the primary Article 50 implementation reference. [74, 75]
AI Vendor/Procurement Standard	Model/SaaS due diligence, contractual controls, concentration and exit.
Privacy/IP Guidance	Personal data, DPIAs, copyright, confidential information and data rights.
Model/Agent Change Standard	Reassessment triggers for model, scope, data, integration or permission changes.
AI Incident Management Playbook	Detection, containment, kill switch, multi-regime notification, investigation and learning; ingest external incident intelligence such as OECD AIM, AI Incident Database and relevant OWASP/vendor exploit round-ups into scenarios and control improvement. [104, 105]
AI Recordkeeping Standard	Logs, evidence, decisions, evaluation artefacts, business records, legal-hold/e-discovery requirements and retention/defensible disposal.
AI FinOps Standard	Budgets, granular cost attribution, quotas, execution budgets, circuit breakers and outcome-level unit economics.

Artefact	Purpose
AI Service Management Standard	Service catalogue, CMDB/configuration, support, SLA/SLO, release and retirement.
AI Literacy and Competency Framework	Training requirements by role.
Workforce Impact & Consultation Standard	Role/job impact, employee-monitoring assessment, employment-law/collective-consultation triggers, worker communication and evidence.
Agent Tool & Protocol Security Standard	Tool registry, metadata integrity, server/version approval, protocol/transport patterns, sandboxing, egress, secrets and change reassessment.
Exception and Waiver Process	Formal deviation with grantor defined by risk tier, no self-approval, scope/rationale, compensating controls, named risk owner, maximum duration/expiry, register entry and periodic reporting to the Executive AI Council for material exceptions. A failed control remains failed/waived; the evidence is not rewritten as passed.

For a European enterprise, these should explicitly map controls to the AI Act rather than maintaining a separate compliance project. The Commission's current implementation guidance confirms that Article 50 transparency obligations have applied since 2 August 2026, including disclosure in relevant human-AI interactions and requirements concerning AI-generated or manipulated content. [16, 8]

Privacy also remains a distinct obligation rather than being superseded by the AI Act. The European Data Protection Board's AI-model opinion addresses, among other matters, when models can be considered anonymous, legal bases for personal-data processing and the consequences of unlawfully processed training data. The EDPB **adopted a harmonized DPIA template on 14 April 2026**; a DPIA remains required where processing is likely to result in high risk to individuals' rights and freedoms. [29, 30]

Organizations in financial services have another overlay: DORA has applied since **17 January 2025** and establishes harmonized requirements for ICT risk management, incidents, resilience testing and third-party risk; financial entities must also maintain information regarding contractual relationships with ICT third-party providers. [31, 32]

Organizations in sectors covered by **NIS2** should likewise treat AI services, agent gateways and critical model/tool dependencies as part of the organization's broader cybersecurity risk-management and supply-chain obligations rather than as a separate innovation environment. Applicability depends on entity and sector scope and national implementation. [60]

The consequence is that the AI governance team should maintain a **regulatory-control crosswalk**, not duplicate governance systems for AI Act, GDPR, cybersecurity, operational resilience and sectoral requirements.

Maintain a live regulatory obligations calendar

For European organizations, "AI Act compliant by 2026" is too imprecise. The current regime contains parallel timelines. The Regulation entered general application on **2 August 2026**, but important obligations phase in on different dates. Article 50 transparency obligations apply from **2 August 2026**; a narrow transition allows providers of systems placed on the market before that date until **2 December 2026** to meet Article 50(2). New Article 5 prohibitions introduced by Regulation (EU) 2026/1744 — covering specified AI-generated/manipulated non-consensual intimate material and child sexual abuse material — apply from **2 December 2026**. Core Annex III high-risk obligations apply from **2 December 2027**, while high-risk systems embedded in Annex I regulated products apply from **2 August 2028**. [8, 35, 36]

This creates several distinct implementation streams rather than one generic compliance project:

1. **Live obligations now:** applicable prohibitions, transparency/disclosure, GPAI/governance duties, privacy, cyber/sectoral requirements and national enforcement readiness.
2. **2 December 2026 changes:** implement the new Article 5 prohibited-use controls and complete the narrow Article 50(2) transition for pre-existing generative systems.
3. **High-risk readiness:** classification, provider/deployer role determination, quality/risk management, technical documentation, logging, human oversight, accuracy/robustness/cybersecurity, conformity,

registration, post-market duties and applicable explanation/contestability routes ahead of the 2027/2028 dates where applicable.

The AI Act is only one part of the obligations calendar. Adjacent EU law should be screened by applicability rather than assumed to apply to every AI use case:

Instrument	Current implementation point	Enterprise relevance
GDPR Article 22 / automated decision-making	Live now. The CJEU's SCHUFA judgment confirms that automated scoring can itself fall within Article 22 where a third party relies strongly on it for a significant decision. [59, 80]	Identify solely automated decisions with legal or similarly significant effects; establish the permitted basis/safeguards and meaningful human intervention where required.
Cyber Resilience Act (EU) 2024/2847	Manufacturer vulnerability/severe-incident reporting applies from 11 September 2026 ; main obligations apply from 11 December 2027 . [76]	Relevant where the enterprise is a manufacturer of a product with digital elements, including software products within scope; add a manufacturer-status check rather than treating all internal AI as CRA-regulated.
Product Liability Directive (EU) 2024/2853	Member-State transposition by 9 December 2026 ; the new regime covers products placed on the market/put into service from that point, and "product" expressly includes software. [77]	Provider/product teams should include defect/safety, update and evidence considerations in product readiness and contracting.
Data Act (EU) 2023/2854	Applies from 12 September 2025 ; Article 3(1) connected-product design duty applies to products/services placed on market after 12 September 2026 ; switching charges are prohibited from 12 January 2027 . [78]	Use cloud/data-processing switching rights and portability requirements to strengthen AI supplier exit design.
AI Liability Directive proposal	The proposal was formally withdrawn in October 2025. [79]	Do not maintain a compliance workstream for a proposal that is no longer live; continue to track applicable national and EU liability law.
Data/Digital Omnibus proposals	Proposed measures remain separate from the enacted Digital Omnibus on AI (Regulation (EU) 2026/1744) .	Keep proposed changes in horizon scanning and do not describe them as current law.

A compact **AI Act implementing-measures register** should sit beside the legal calendar because operational readiness increasingly depends on Commission guidance and standards status as well as the Regulation itself:

Implementing instrument	Status at 22 September 2026	How the enterprise should use it
Code of Practice on Transparency of AI-generated Content	Final 10 June 2026; assessed adequate by the Commission/AI Board in July 2026. [74]	Primary implementation reference for relevant Article 50(2), (4) and (5) marking/labelling commitments; voluntary adherence is not conclusive proof of compliance.
Article 50 transparency guidelines	Final 20 July 2026. [75]	Interpret scope, exemptions and provider/deployer transparency duties; use with the Code rather than relying on C2PA or a vendor scheme alone.
High-risk classification guidelines	Draft published May 2026; consultation closed July 2026. [108]	Useful Commission interpretation, but label as draft until final adoption.
GPAI guidance / Code of Practice	Current Commission guidance and voluntary Code apply to GPAI obligations. [62, 111]	Determine whether downstream fine-tuning/modification changes provider status/obligations; record provider

Implementing instrument	Status at 22 September 2026	How the enterprise should use it
Harmonised standards	Track the Commission/CEN-CENELEC standardisation programme and verify Official Journal citation status before relying on a standard for presumption of conformity. [109, 110]	documentation and Code reliance. Maintain standards status in the obligations register; published European standards are not automatically OJ-cited harmonised standards.
Grandfathering / sandbox	Article 111 preserves defined transition rules for pre-existing systems; public-authority high-risk systems have a 2 August 2030 long-stop. The national sandbox deadline is 2 August 2027 . [35, 36, 83]	Add grandfathering and sandbox availability to inventory/readiness decisions rather than assuming all systems share one transition date.

Automated decision-making, meaningful human involvement and contestability

The **GDPR Article 22** layer applies independently of the later AI Act high-risk timetable. The inventory/intake process should identify decisions based solely on automated processing that produce legal or similarly significant effects, including cases where an automated score is determinative in practice. SCHUFA confirms that a scoring output can itself constitute automated individual decision-making where a third party draws strongly on it to establish, implement or terminate a contractual relationship. [59, 80]

Where an exception permitting such processing applies, design the safeguards as an operating process rather than a notice: appropriate information, a route to obtain **meaningful human intervention**, the ability to express a view and contest the outcome, and evidence that the reviewer has competence, authority, time and information to change the result rather than rubber-stamp it. EDPS and Dutch AP guidance provide practical tests for meaningful intervention. For high-risk AI, Article 14(4)(b) specifically requires human overseers to remain aware of automation bias; Article 26(2) requires deployers to assign oversight to people with the necessary competence, training and authority; and Article 86 creates explanation rights for defined individual decisions based on high-risk AI output. [35, 81, 82]

Determine the enterprise's AI Act role before applying duties

The organization should determine its legal **operator role per system and per activity**, not label itself globally as a "deployer" or "provider." A business may hold multiple roles across its AI estate. The intake gate should ask:

1. **Did we develop or have the system developed and place it on the market or put it into service under our own name/trademark?** If yes, assess provider duties.
2. **Are we using a third-party system under our authority?** If yes, assess deployer duties.
3. **Are we importing or distributing a system into/on the Union market?** If yes, assess importer/distributor duties.
4. **Have we changed the system or its intended purpose?** Under Article 25, an importer, distributor, deployer or other third party can become the **provider of a high-risk AI system** if it puts its own name/trademark on the system, makes a substantial modification while it remains high-risk, or changes the intended purpose so a previously non-high-risk system becomes high-risk. [35]

For a high-risk system where the enterprise is the provider, the readiness checklist should cover the applicable Chapter III provider obligations, including conformity assessment/CE marking where relevant, EU-database registration for Annex III systems, technical/risk/quality documentation, post-market monitoring and serious-incident reporting. Public-authority deployers have separate registration duties. [35]

Distinguish FRIA from DPIA and generic impact assessment

Article 27 Fundamental Rights Impact Assessment (FRIA) should be a named legal artefact with an applicability decision, not buried inside the generic AI-system impact assessment. Its scope is narrower than "all employment AI": before first use of an Article 6(2) high-risk system (other than Annex III point 2), FRIA applies to deployers that are bodies governed by public law or private entities providing public services, and it also applies to deployers of Annex III 5(b) creditworthiness and 5(c) life/health-insurance systems. It is therefore **not automatically triggered for an**

ordinary private employer solely because a system falls within Annex III employment point 4. Where applicable, the FRIA can cross-reference relevant DPIA material under the amended Article 27, but it remains a distinct obligation and notification artefact. [35, 57]

Article 50 requires relevant providers to make certain synthetic-content outputs machine-readable and detectable, and deployers have disclosure duties in defined cases. The Commission's **Code of Practice on Transparency of AI-generated Content** (final 10 June 2026 and subsequently assessed adequate) and the **Article 50 Guidelines** (20 July 2026) now provide the primary EU implementation reference. The enterprise should establish a **synthetic-content transparency pipeline** spanning generation metadata, disclosure logic, provenance where appropriate, retention of evidence and downstream publishing controls. C2PA Content Credentials are one potentially useful technical means of implementing machine-readable provenance/markings — the 2.4 specification includes an AI disclosure assertion — but neither C2PA nor any particular watermarking product is mandated by the law. [74, 75, 51]

National implementation also matters. In Sweden, the Government assigned PTS, IMY, Finansinspektionen, Läkemedelsverket and Swedac to act as national competent authorities in June 2026, but the **current assignment is interim and runs through 31 December 2026** while supplementary national legislation is prepared. SOU 2025:101 was delivered in October 2025 and the consultation stage followed; PTS describes the 2026 authority assignments as operating **in anticipation of complementary Swedish legislation**. Sweden's national regulatory sandbox is still being prepared and must be operational by **2 August 2027**. [37, 83, 84, 85] Enterprises should therefore verify the current national allocation and sandbox status before relying on them.

Workplace AI has a separate future high-risk obligation that should not be confused with Article 50 or FRIA. **Article 26(7)** requires employer deployers, before putting into service or using a high-risk AI system at the workplace, to inform workers' representatives and affected workers that they will be subject to its use. Because Article 26 is within the delayed high-risk obligations, this duty follows the relevant **2 December 2027 / 2 August 2028** application dates depending on classification. Existing national information/consultation duties can arise earlier and independently. [35, 36]

Orchestrate incident reporting across regimes

One AI incident can trigger several legal/regulatory clocks. The AI Incident Management Playbook should therefore maintain a **single notification-orchestration record** containing the incident start/awareness time, applicable regimes, next deadline, accountable owner, authority/recipient and evidence submitted.

Regime	Core trigger / timing to operationalize
EU AI Act	Providers of high-risk AI systems report serious incidents generally to the competent market-surveillance authority, with the statutory timing varying by severity; for systems under the AI Office's exclusive competence under Article 75, the reporting route is to the AI Office. Deployers have separate immediate-information duties when they identify a serious incident. [35]
GDPR	Where a personal-data breach is not unlikely to result in risk to individuals, the controller notifies the supervisory authority without undue delay and, where feasible, within 72 hours of awareness. [59]
NIS2	For in-scope significant incidents: 24-hour early warning, 72-hour incident notification and final report generally within one month. [60]
DORA / sector regimes	Financial and other regulated entities should map the applicable sector incident taxonomy, forms and deadlines to the same orchestration record rather than assume the AI Act clock replaces them. [31, 32]
Cyber Resilience Act	Where the enterprise is an in-scope manufacturer, actively exploited vulnerabilities and severe incidents have reporting obligations from 11 September 2026. Route these through the same incident record rather than treating product-security reporting as separate from AI operations. [76]

The purpose is not to make one team expert in every regime; it is to prevent a serious AI incident from being handled technically while a parallel legal notification clock is missed.

Screen prohibited enterprise and workplace uses explicitly

"Check prohibited AI" is too abstract for procurement and HR intake. At minimum, the screening catalogue should name the Article 5 categories most likely to appear in enterprise products: manipulative/exploitative techniques within the statutory tests; social scoring; **emotion inference in workplaces and education institutions** except where the medical/safety exception applies; biometric categorisation that infers the protected/sensitive characteristics covered by Article 5; and other prohibited biometric or predictive-policing uses relevant to the estate. From **2 December 2026**, the new Article 5(1)(ba)/(bb) prohibitions concerning defined non-consensual intimate material and child sexual abuse material also apply. [35, 36]

Article 50(3) is different from prohibition: where an allowed emotion-recognition or biometric-categorisation system is used, deployers have transparency/information duties to exposed natural persons. Keep **prohibition screening, transparency and high-risk classification as separate questions**. [35, 75]

5.3 Security and resilience

Security is a no-compromise boundary condition across every maturity stage. The objective is not a parallel AI-security bureaucracy; it is to extend the existing cyber, IAM, secure-development and incident-management system with the AI-specific attack surfaces and autonomous authorities that conventional controls do not already cover.

Agent security and non-human identity

Security should extend existing enterprise controls rather than inventing an AI-only security universe. Google's Secure AI Framework emphasizes inventories of models, datasets and AI artefacts, access control, data integrity, supply-chain security and AI-specific threat testing. The AIMS should reuse enterprise ISMS and secure-development baselines — for example ISO/IEC 27001/27002 and the NCSC/CISA secure-AI guidance — and add AI-specific threat/evaluation taxonomies such as MITRE ATLAS and NIST AI 100-2 where useful. ETSI TS 104 223 and the joint NSA/CISA AI Data Security guidance are useful additional control references where applicable; the objective is integration, not a parallel cyber programme. [33, 87, 88, 89, 90, 91]

For agents, however, a genuinely new asset class appears: **the non-human autonomous actor**. NIST began a dedicated AI Agent Standards Initiative in 2026 and has identified agent identity, authorization, auditability, non-repudiation and prompt-injection resistance as emerging control problems. [27, 28]

Consequently, every production agent should have:

a named business owner + a unique machine identity + an explicitly defined purpose + approved data + approved tools + task-scoped privileges + defined action limits + evaluation evidence + runtime telemetry + lifecycle status + an expiration/review date + a rapid disable mechanism.

Microsoft's current agent-security guidance likewise recommends unique agent identities, least-privilege RBAC, task-scoped permissions, action allowlists, approvals or time-bound elevation for privileged actions, full auditability and revocation. [14, 15]

That principle should become as normal as identity management for human employees. The same zero-trust logic applies when **external agents or machine clients interact with enterprise services**: distinguish machine principals from humans where material, apply rate/action limits, expose only agent-safe capabilities, log machine-originated activity and do not assume that a third-party agent inherits all authority of the person or organisation it represents. [61]

A useful threat-model screening question is whether the design combines **private data + untrusted content + an exfiltration/action channel**. Where all three are present, architectural separation and deterministic mediation should be assumed necessary rather than relying on prompt instructions or a classifier alone. [87]

Agent identity, delegation and tool/protocol supply-chain security

NIST's 2026 agent identity and authorization concept paper remains an **initial public draft** (its comment period closed on 2 April 2026 and no final version had been published by September 2026), but its direction is important: identification, authorization, auditing, non-repudiation and prompt-injection resistance become first-class design

questions when software agents receive access to enterprise tools and data. [41] The enterprise should therefore be able to reconstruct **which agent acted, which human or service delegated authority, what purpose and scope applied, which tool and version was used, what data crossed the boundary, and what deterministic policy permitted the action**. This can be treated operationally as an authority/delegation provenance requirement, while being clear that the terminology is an enterprise synthesis rather than a finalized NIST standard.

Every agent needs an accountable identity.



Reference pattern for non-human identity and delegated authority

NIST does not mandate one implementation stack. A practical enterprise pattern is nevertheless possible:

Lifecycle step	Enterprise control objective	Example implementation patterns
Register	Every agent/service identity has a named owner, purpose, environment, risk tier and expiry/review date in IAM/IGA and the AI register.	Non-human identity record linked to service/CMDB/AI inventory.
Authenticate	The running workload proves its identity using short-lived, automatically rotated credentials rather than embedded long-lived secrets.	Workload identity such as SPIFFE/SPIRE is one standards-based option. [54]
Delegate	Human/service authority is transformed into purpose-, audience-, scope- and time-bounded authority for the agent rather than copied wholesale.	OAuth 2.0 Token Exchange supports impersonation/delegation patterns; use only where it fits the enterprise IAM architecture. [55]
Authorize	A deterministic policy decision point evaluates identity, resource, purpose, scope, transaction limits, segregation-of-duties rules and risk context before action.	Enterprise policy engine / authorization service; the LLM does not self-authorize.
Enforce	Policy is enforced at the API/tool boundary so disallowed actions are technically impossible, not merely discouraged by prompts.	Gateway, API proxy, service mesh or tool broker as policy enforcement point.
Audit	The enterprise can reconstruct the delegation chain, policy decision, tool/version, action and outcome.	Tamper-resistant trace/event records linked to identity and transaction IDs.
Attest	Agent access and delegated rights are periodically re-certified and reassessed after material change.	IGA access review plus AI change-governance trigger.

Lifecycle step	Enterprise control objective	Example implementation patterns
Revoke	Identity, credential, delegation and tool access can be disabled quickly and independently.	Central revocation/kill path tested as part of operational acceptance.

SPIFFE/SPIRE and RFC 8693 are **examples, not mandated technologies**. The architectural requirement is a standards-based identity and delegation chain that is short-lived, least-privilege, independently enforceable and auditable. [41, 54, 55]

Protocols such as the Model Context Protocol (MCP) make tool integration easier but do not remove the need to secure the tool supply chain. OWASP's Agentic Top 10 identifies goal hijacking, tool misuse, identity/privilege abuse, supply-chain risk, unexpected code execution and insecure inter-agent interactions as distinct agentic risk classes. [5] Peer-reviewed MCPTox research provides concrete evidence that malicious instructions embedded in tool metadata can manipulate tool-using agents; its test set covered 45 live MCP servers and found substantial attack success across evaluated agents. [42]

The architectural response should be protocol-neutral even where MCP is currently prominent:

- maintain an **approved tool/server registry** with named ownership, purpose and risk classification;
- inspect tool descriptions, schemas and metadata before registration and treat metadata as untrusted input;
- pin or cryptographically verify approved versions/configurations where feasible, and trigger re-approval on material change;
- separate discovery from execution and enforce server/tool allowlists at runtime;
- constrain outbound network access, secrets and filesystem access;
- sandbox local tool processes and agent execution environments;
- enforce action-level authorization and transaction limits outside the model;
- record tool calls, inputs/outputs, policy decisions, versions and failure events;
- provide rapid revocation of a tool, agent identity or delegated permission.

Implementation detail matters. The **2026-07-28 MCP specification** moved the protocol to a stateless core for remote HTTP use, added header-based routing and authorization hardening, and reclassified the legacy HTTP+SSE transport (deprecated since the 2025-03-26 version) under the feature-lifecycle policy with a minimum twelve-month deprecation window; no fixed removal date had been set. [52, 43] The stable Enterprise-Managed Authorization extension also provides an enterprise pattern for central identity-provider control over MCP access. [53]

For enterprise remote-tool traffic, the default pattern should therefore be **brokered/gatewayed access**: approved servers and tools, identity-aware authorization, policy enforcement, egress controls, request/tool metering, version/configuration evidence and centralized logging. The gateway can be an MCP-aware component or a broader enterprise API/tool control plane; the control objective matters more than the product label.

Security research published by OX in April 2026 documented command-injection/RCE patterns in implementations that allowed untrusted MCP configuration to reach local STDIO process spawning, with OX estimating exposure across 150M+ downloads and up to 200,000 vulnerable instances at the time of disclosure. The initial disclosure reported **10 CVEs issued and counting**, not the 40+ figure sometimes repeated in secondary commentary. [44, 112]

That evidence still does **not** justify a universal ban on STDIO. Local STDIO can be acceptable when configuration is trusted and not user-controlled, executables/arguments are constrained, the process is sandboxed with least privilege, secrets/filesystem/network access are bounded, and the host trust boundary is appropriate. Where those conditions cannot be demonstrated, use an isolated remote/brokered pattern instead.

Agentic software engineering and infrastructure agents

AI coding, infrastructure and CI/CD agents should be governed as **production agents whenever they can change source, build artefacts, pipelines, infrastructure-as-code, cloud resources or deployment configuration**. They are not lower risk merely because they live in an IDE or terminal. Their effective tools can include the filesystem, shell, package manager, source-control credentials, cloud CLIs and deployment APIs.

Minimum controls should include: a distinct machine/workload identity where practical; short-lived scoped credentials; **no standing production credentials in ordinary developer-agent sessions**; isolated/sandboxed execution; constrained command and network access; protected branches; mandatory CI and human/independent review for production paths; separation of code generation from production deployment; secrets isolation; auditable tool/command traces; and tested revocation. Package selection should use approved/private registries or other dependency controls where material, because hallucinated or attacker-controlled package names can turn code generation into a supply-chain path. Maintain SBOM and, where models/datasets/AI components matter to the release, ML-BOM/model-provenance evidence. [95, 96]

Documented 2025-2026 incidents and security advisories involving coding assistants, developer extensions and package/tool chains demonstrate why this class needs explicit treatment rather than being hidden inside generic "developer tooling". The lesson is not that coding agents are inherently unsafe; it is that **developer convenience can collapse segregation between probabilistic planning and privileged execution**. [97, 98, 99, 100]

UI-acting and computer-use agents

Agents that operate through browsers, desktops or RPA-style user interfaces should **not run inside a human's normal authenticated session for consequential activity**. Prefer an isolated browser/desktop or workload context with its own identity, approved domains/applications, action classes, session recording and explicit confirmation for sensitive actions. Where a policy-enforced API or equivalent boundary cannot be introduced, treat the agent's effective authority as potentially **unbounded by the AI control plane** and cap it at *Prepare action* unless a stronger assurance case demonstrates containment.

The design test is simple: **can every consequential action be intercepted, attributed and denied outside the model?** If not, the enterprise should reduce authority, isolate the execution surface or redesign the integration. [15, 49]

Multi-provider architecture should also include explicit **deprecation and continuity controls**: approved model versions; provider/end-point region constraints; advance assessment of announced model retirement; tested fallback to another approved model or a deterministic/manual path; and a rule that failover must not silently widen data access, geography, permissions or risk.

Continuity testing should include at least one **correlated-failure scenario**, not only a single-provider outage: for example, a preferred model is deprecated while a second provider becomes unavailable in an approved region or can no longer meet a residency/contractual constraint. The fallback decision must preserve the original data classification, authority, geography and control envelope; otherwise the service should degrade to a safer deterministic/manual path rather than silently fail open.

For a lower-level catalogue of agent control mechanisms, implementation patterns, testing ideas and evidence expectations, cross-reference: Cognostrix, "Agentic AI Controls, Governance and Assurance — A Reference Catalogue", v1.0 (September 2026).

5.4 Platform and architecture

Platform and engineering excellence are related but not identical. The platform is the shared product/service layer — model access, gateways, reusable AI services, runtime infrastructure and observability. Engineering excellence is the delivery discipline used to design, build, test, release and operate AI-enabled products on or around that platform.

The technical control plane

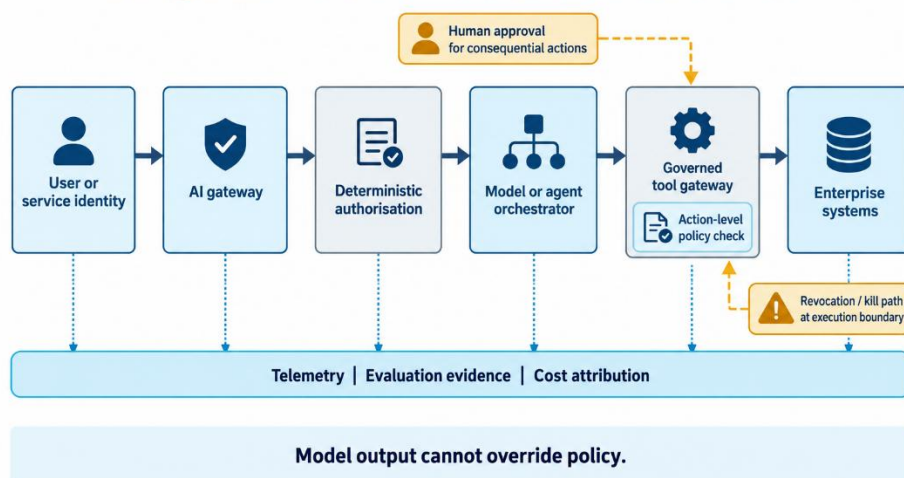
The platform should support **multiple AI technologies and providers rather than becoming synonymous with one LLM vendor**.

At maturity, the common foundation should include:

Layer	Enterprise capability
Model layer	Approved model catalogue across commercial, open and internally developed models; version control; model selection/routing.

Layer	Enterprise capability
AI gateway	Authentication, policy enforcement, logging, rate limits, provider abstraction, model routing, request-level metering and cost accounting.
Identity layer	SSO/RBAC for users; non-human identity lifecycle for agents/services; short-lived credentials; ownership, attestation and scoped delegated authorization.
Authorization/policy plane	Deterministic policy decision and enforcement points for purpose, scope, read/write rights, segregation of duties, transaction limits, egress, budget and high-risk action approval; model output cannot override policy.
Data layer	Classification, catalogue, lineage, quality, access control, governed structured and unstructured data.
Knowledge/RAG layer	Reusable document ingestion, indexing, retrieval, citation/provenance and access-aware search, with lineage, retention and deletion propagation across source, chunk, index, embedding and cache layers.
Tool and integration layer	Governed API/tool registry, connectors, tool permissions, transaction constraints and secrets management; approved protocol patterns; metadata integrity; version pinning/change control; and runtime egress/tool-call policy.
Development layer	Sandboxes; environment separation for prompts, agent configuration and secrets; SDKs; reusable components; protected CI/CD and controlled promotion.
Evaluation layer	Functional quality, groundedness, hallucination, safety, security, fairness, robustness, regression, latency and cost tests.
Runtime guardrail layer	Deterministic boundaries: IAM/policy checks, allowlists, schema/action validation, DLP enforcement, approval gates, execution budgets, rate/recursion limits and circuit breakers. Probabilistic risk-reduction controls: classifiers, model-based safety checks and judge models, with measured detection/false-positive performance.
Observability layer	Model/agent telemetry, traces, tool calls, decisions, state transitions, cost, token/context growth, incidents and business outcomes; prefer vendor-neutral telemetry patterns where practical. OpenTelemetry GenAI semantic conventions are an emerging direction and remain Development-status rather than a settled standard. [103]
Operations layer	MLOps, LLMOps and AgentOps together with existing DevSecOps and ITSM; provider/model deprecation watch, failover and controlled fallback paths; defined recovery objectives for material vector indexes, memory/state stores and agent configuration.

Intelligence inside enforceable boundaries



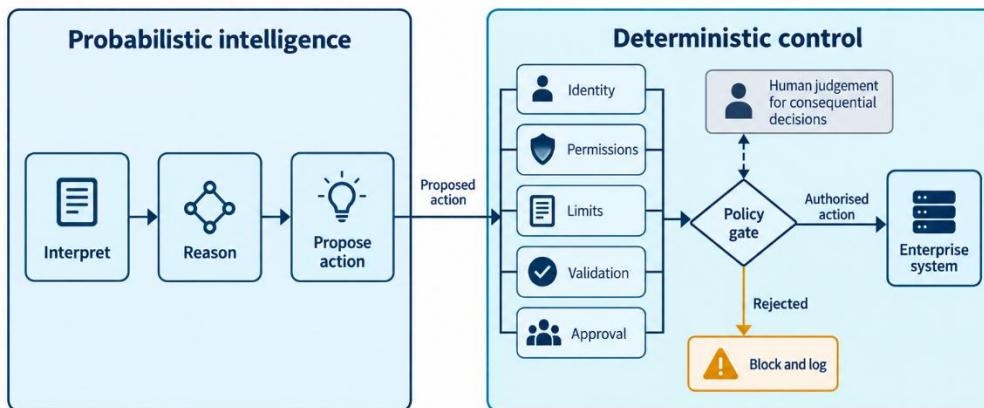
The data foundation deserves particular attention. Recent scaling research identifies data limitations as a major barrier to agentic deployment and argues for moving beyond use-case-specific data pipelines toward reusable, governed data products, including stronger treatment of unstructured information. [19, 20]

The reference control flow should be understandable without relying on any one vendor stack:

User/service identity → AI gateway → deterministic authorization/policy decision → model or agent orchestrator → governed tool/MCP gateway → enterprise systems, with telemetry, evaluation evidence and cost attribution captured across every material hop.

Boundary assumption: every deterministic control in that flow assumes the consequential action traverses an API, tool broker, gateway or other enforceable policy point. Some agent classes do not. A browser/computer-use agent can click through a human-authenticated session; an IDE agent can execute local commands; an infrastructure agent can mutate cloud resources from a developer shell. If those execution surfaces are not separately controlled, the model can inherit more authority than the formal agent record suggests.

Models propose; enforced policy permits.



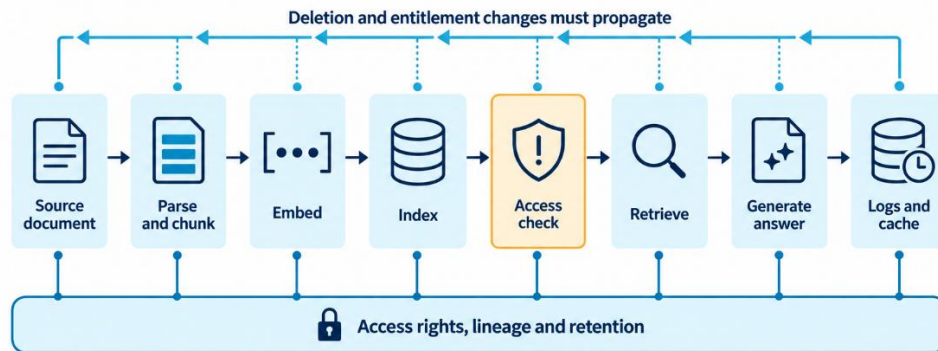
5.5 Data and knowledge

Privacy and lifecycle controls for RAG and vector systems

A governed RAG layer is not merely a search feature. It creates derived representations, indexes and retrieval paths that can expose information differently from the source repository. The EDPB's position on AI models is deliberately contextual: whether an artefact can be regarded as anonymous depends on whether identification or extraction of personal data is sufficiently unlikely in the circumstances, assessed case by case. [39] Research on RAG membership inference further demonstrates that attackers may be able to infer whether a target document is present in a retrieval database from system behaviour, including in black-box settings. That evidence should be treated as an emerging security threat, not as a statement that every embedding is automatically personal data. [40]

For material RAG services, the data lifecycle should therefore cover **source document → parsed/chunked representation → embedding/index → caches and intermediate stores → retrieved context → model or agent output**. Access control should be enforced at retrieval time, not assumed from the source repository alone. Retention and erasure processes should identify and remove relevant derived artefacts where required; ingestion jobs should preserve source lineage; and high-risk corpora should be tested for retrieval leakage, cross-tenant exposure and prompt-based extraction.

Govern the derived knowledge, not only the source.



The practical control objective is not "encrypt the vector database". It is: **a user or agent should never gain a retrieval path to information they were not authorized to obtain, and deletion or access changes should propagate through every material derived store.**

For production services, make that objective measurable. Define **access-change and deletion-propagation SLOs**, test them periodically, and retain evidence that a revoked entitlement or erased source item no longer appears through retrieval, caches or derived indexes within the approved window. Apply normal enterprise encryption requirements to relevant stores and transport, but do not imply that a special form of "embedding encryption" is a universal GDPR requirement.

RAG governance must also answer **who owns the retrieved content and how current it must be**. Material corpora should have named content owners, authoritative-source rules, version/conflict handling and a freshness SLO appropriate to the decision. Retrieval evaluation should include stale-document and conflicting-version cases, not only relevance scores.

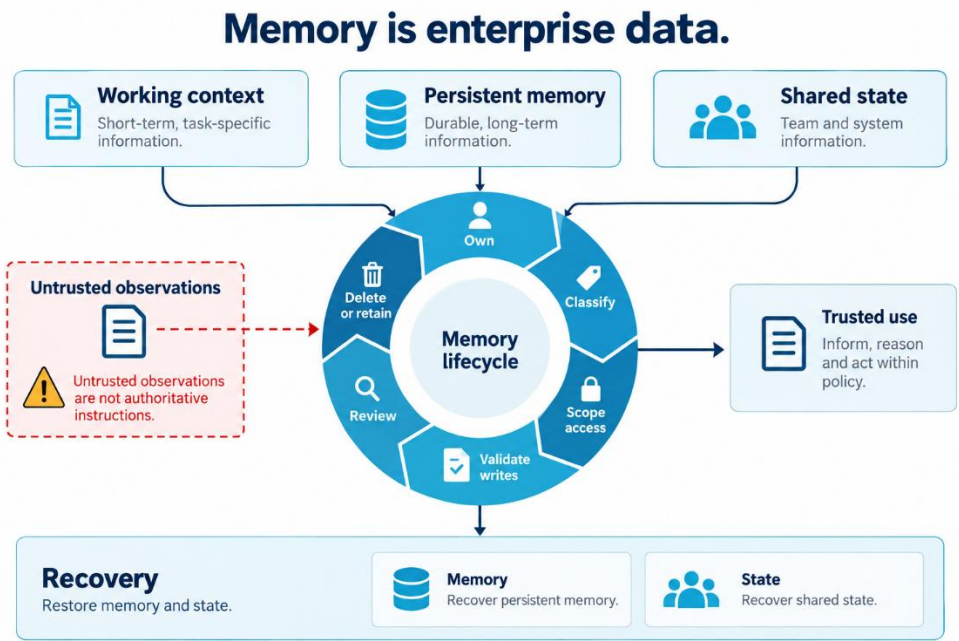
Govern agent memory and persistent state as enterprise data

Agent memory is not merely a convenience feature. Persistent state can contain personal data, confidential information, inferred facts, poisoned instructions or stale decisions, and it can cross agent boundaries if not deliberately contained.

Every material memory/state store should therefore define:

- **scope** — per user, per case, per agent, shared team or enterprise;
- **data classification and lawful purpose** — what may be retained and why;
- **write authority** — which actors or agents may create, amend or promote memory;
- **provenance** — source, timestamp, author/agent and confidence/status where relevant;
- **retention/TTL** — default expiry and conditions for longer retention;
- **sharing boundaries** — which agents, tools or domains may read it;
- **deletion and correction propagation** — including downstream caches or summaries where material;
- **poisoning/staleness controls** — validation, quarantining and revalidation of high-impact retained state; and
- **rollback/recovery** — the ability to restore a known-good state after corruption or faulty autonomous action.

Memory controls should be proportionate to persistence and consequence: ephemeral conversation context is not governed identically to a long-lived shared operational memory that can drive transactions. [5, 45, 46]



For production acceptance, convert those objectives into repeatable tests. Useful patterns include: seed a controlled document or memory item; verify retrieval only for entitled identities; revoke access or erase the source; wait no longer than the approved propagation window; re-query source, chunks, indexes, caches, summaries and persistent agent state; and retain the negative-result evidence. For high-impact persistent memory, also inject a known poisoned/stale state entry in a test environment and demonstrate that validation/quarantine prevents it from becoming authoritative or driving an irreversible action.

5.6 Delivery and engineering

AI delivery should extend the organisation's existing product, SDLC, DevSecOps and change practices. The AI-specific uplift is the addition of model/data/prompt/agent artefacts, evaluation evidence, probabilistic-system testing, model/tool version triggers and autonomy-aware release controls.

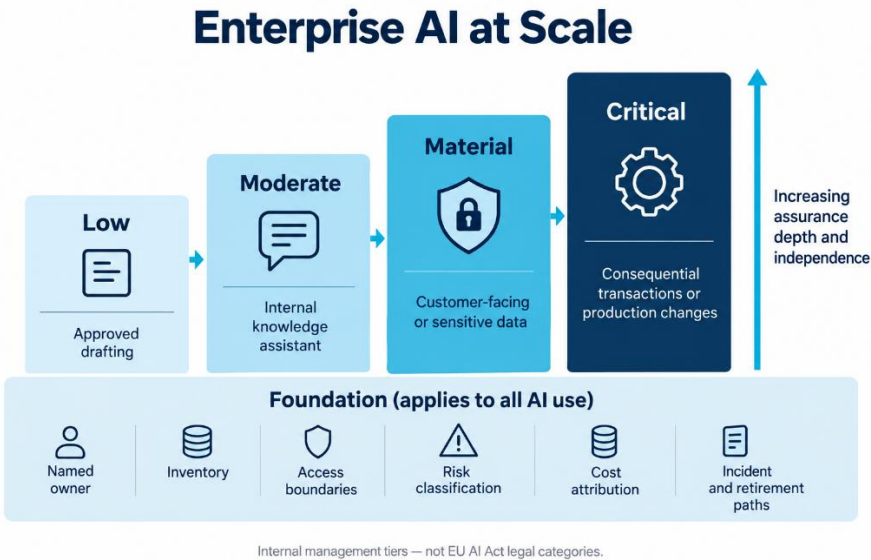
Risk-tier the lifecycle

An internal enterprise classification might look like this:

Internal tier	Typical example	Governance
Low	Employee drafting/summarization using approved nonsensitive data	Approved platform, training and baseline controls.
Moderate	Internal knowledge assistant or operational recommendation	Registered use case, evaluation, owner, data/security review.
Material	Customer-facing GenAI or significant decisions/sensitive data	Formal risk review, stronger testing, monitoring, human oversight where appropriate, and content-safety/consumer-harm evaluation for customer-facing generation.
Critical	High-impact decision; agent executing consequential transactions; coding/infrastructure agent with production-change authority	Executive/risk approval, stringent assurance, constrained authority, protected deployment path, continuous monitoring and independent review.

This is an **internal management system** and must not be confused with the legal risk categories under the EU AI Act.

The lifecycle should be proportionate: a meeting-summary tool should not undergo the same assurance process as an employment-screening model.



Make permissions and scope changes governance events

One of the most useful requirements in the job-specification corpus concerns **re-assessment when an integration changes scope**—for example when a connector moves from read permissions to write permissions. It requires a revised technical/security assessment, approvals and controlled rollout rather than treating the existing approval as permanent.

This should become enterprise policy.

Use the consolidated **Material-change trigger list** in the practitioner artefacts below. At minimum, increases in permission scope, write/action authority, autonomy, model/provider/version, data/RAG/memory, population or purpose, geography, external exposure or downstream integrations should force reassessment.

An AI approval is **approval of a defined system configuration and purpose**, not a permanent license to evolve without review.

5.7 Systemic assurance and human oversight

Introduce an explicit agent autonomy ladder

Agents are important enough to require a separate authority model.

Authority	What AI may do	Default control
Assist	Draft, retrieve, summarize	Human decides and acts.
Recommend	Analyze options and propose decisions	Human authorizes decision.
Prepare action	Populate transaction or prepare workflow	Human confirms execution.
Execute bounded action	Perform predefined reversible actions	Limits, monitoring and exception rules.
Orchestrate	Select among tools and complete a bounded multi-step workflow	Task-scoped permissions, approvals for sensitive actions.
High autonomy	Operate independently inside a narrow controlled domain	Exceptional assurance case, transaction limits, auditability, rapid shutdown.



The default principle should be:

No action authority by default; expand authority only when business value justifies it and evidence demonstrates that the risk is controlled.

That is consistent with Microsoft's current agentic-security guidance, which recommends least privilege and "least action," deterministic controls, explicit approval for high-risk actions and visible shutdown mechanisms. [15]

OWASP's 2026 Agentic Top 10 underscores why these controls matter. Its threat categories include agent goal hijacking, tool misuse, identity and privilege abuse, memory/context poisoning, insecure inter-agent communication, cascading failures and human-agent trust exploitation. [5, 23]

The **OWASP GenAI LLM Top 10 2026** was published in August 2026 and announced alongside the Agent Control Standard on 1 September 2026. It is grounded in real-world incident evidence and reinforces that the security problem increasingly shifts from **what a model can say to what an AI-enabled system is permitted to do**. [24, 25, 61]

Therefore:

Prompt instructions are not security controls.

This leads to a broader architectural rule: **do not ask a probabilistic model to enforce a control that can be enforced deterministically**. Permission checks, transaction ceilings, separation of duties, schema validity, mandatory fields, rate limits, budget ceilings, executable allowlists, network boundaries and kill conditions should normally be implemented outside the model. Models may recommend; policy engines, code and identity systems decide whether an action is technically possible.

An instruction such as "never delete files" may reinforce intended behavior, but deletion must also be impossible unless a deterministic authorization mechanism grants the agent the relevant action.

For assurance purposes, classify each guardrail. **Only deterministic controls should be counted as hard authority/security boundaries**. Probabilistic classifiers, prompt instructions and LLM-as-judge checks can materially reduce risk, but their detection, false-positive/false-negative and variance characteristics must be measured and they should not be presented as making a prohibited action technically impossible.

Preserve mature model-risk management for predictive and traditional ML

A unified AI management system should not weaken established model-risk disciplines merely because attention has shifted to GenAI and agents. For **material predictive/statistical ML models**, the lifecycle should include an authoritative inventory, materiality/risk tier, documented development and limitations, independent challenge or

validation proportionate to risk, implementation verification, performance/drift monitoring, change control and periodic revalidation. Challenger models, back-testing, benchmarking and outcome analysis should be used where appropriate to the model and decision context rather than prescribed mechanically.

This is especially important in regulated financial services, where supervisory model-risk frameworks already expect model identification/classification, governance, sound development/use, independent validation and model-risk mitigation. The PRA's SS1/23 is one current example and explicitly covers AI/ML techniques to the extent they are used as models; other jurisdictions have their own supervisory frameworks. [68] Outside regulated sectors, the same principles are useful good practice but should be scaled to materiality rather than imported as bank-grade bureaucracy.

Assure multi-agent systems as systems, not collections of individually tested agents

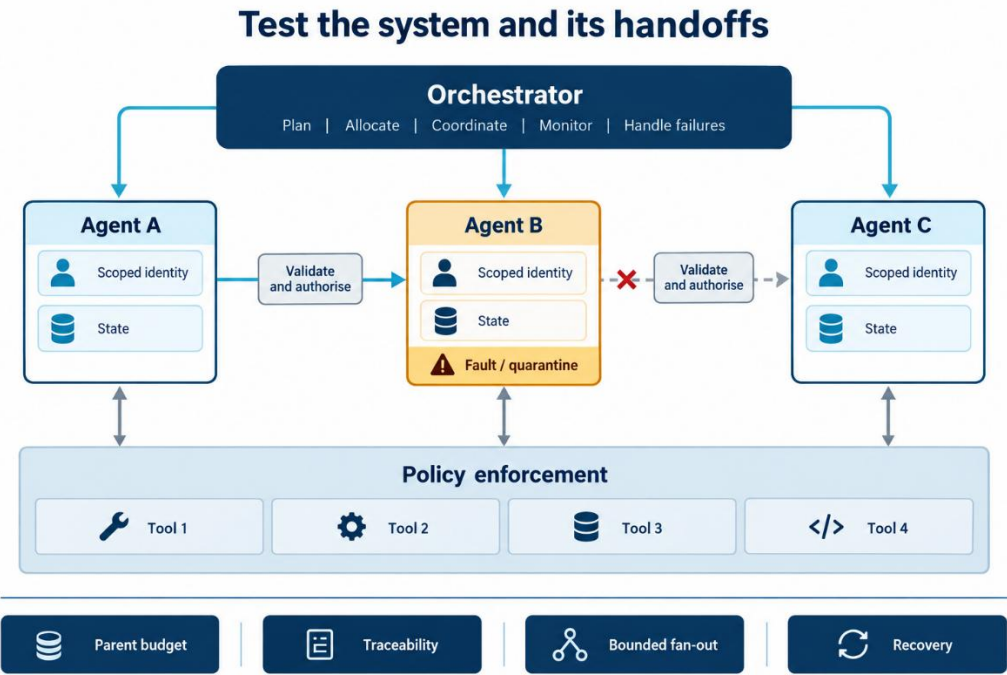
Unit-level evaluation remains necessary but is not sufficient when multiple agents share state, delegate work or act through an orchestrator. Emerging research demonstrates classes of failure that arise from interaction: indirect prompt injection traversing an orchestrator can compromise several agents even where access controls exist, and shared context can propagate corrupted or misleading information across the workflow. [45] Broader research on interacting agents similarly argues that network effects introduce security problems beyond the single-agent case. [46]

The enterprise assurance case should therefore test the **topology and handoffs**, not merely the component prompts. Required patterns include:

- separate or purpose-scoped memory/state where one agent should not be able to overwrite another's authoritative context;
- typed/schema-validated messages between agents;
- explicit identity and authorization at agent-to-agent and agent-to-tool boundaries;
- deterministic validators before irreversible or regulated actions;
- adversarial tests for poisoned shared state, compromised tools, stale data, loops and conflicting instructions;
- end-to-end replayable traces that show delegation, state changes, tool calls and policy decisions;
- bounded concurrency, retries, recursion and fan-out;
- fault containment so one agent's failure cannot automatically acquire the blast radius of the full system.

For a material multi-agent workflow, the release gate should require concrete **system-level exit evidence**:

Exit criterion	Example acceptance evidence
Traceability	Material decisions/actions can be replayed across agent identity, delegation, state transition, model/tool call and policy decision.
Authorization integrity	Negative tests demonstrate that an agent/sub-agent cannot exceed delegated purpose, scope, read/write rights or transaction limits.
State integrity	Poisoned, stale or malformed shared state is rejected/quarantined and does not silently become authoritative.
Fault containment	Failure or compromise of one agent remains inside the defined blast radius and does not automatically propagate privileged action.
Bounded execution	Concurrency, recursion, retry depth, fan-out, elapsed time and parent/child execution budgets remain within configured limits.
Recovery	Kill/revoke and rollback procedures restore a known-good configuration/state and prevent unfinished actions from continuing.
Assurance closure	No unresolved finding above the organization's documented production risk-acceptance threshold; residual risks have named owners.



AgentCheck is an example of the emerging engineering direction: reproduce tool faults, intervene with controlled perturbations and verify whether a mitigation actually removes the failure. [47] The specific tooling will evolve; the enduring control objective is **reproducible failure testing at the system boundary**.

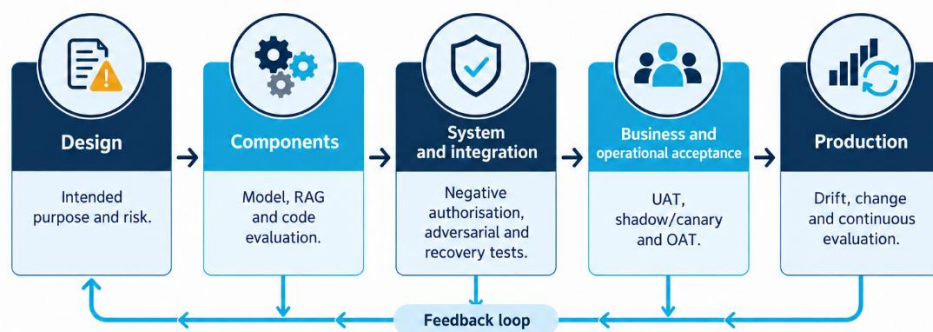
Map traditional assurance stages onto AI and agentic releases

AI does not make traditional testing stages obsolete; it changes what must be tested inside them. ISO/IEC TS 42119-2:2025 explicitly applies established software-testing processes to AI systems using a risk-based approach. [58] A practical enterprise mapping is:

Assurance stage	AI/agentic interpretation	Typical exit evidence
Requirements / design assurance	Intended purpose, prohibited uses, risk tier, data/tool authority, human responsibility, acceptance metrics and impact assessment.	Approved requirements, architecture/threat model, risk/impact record.
Unit / component test	Deterministic code, prompts/configuration, model wrappers, tools/functions, policy rules and parsers tested in isolation.	Automated tests, static checks, prompt/tool regression cases.
Component / model / RAG evaluation	Task quality, groundedness/factuality, retrieval-only quality , safety, bias where relevant, robustness, latency and cost; check evaluation-set contamination/leakage and run-to-run variance.	Versioned evaluation dataset/lineage, retrieval metrics and threshold results.
System test	End-to-end agent/workflow behavior, memory/state, fallback, loops, approvals and business rules.	End-to-end traces and failure/recovery evidence.
System integration test	IAM, APIs, scopes, tool/MCP servers, event flows, data boundaries, observability and negative authorization cases.	Integration test evidence including denied-action tests.
Adversarial / red-team test	Prompt injection, goal hijack, poisoned retrieval/tool metadata, exfiltration, privilege escalation, resource abuse and multi-agent propagation.	Findings, risk acceptance/remediation and retest.
UAT / business acceptance	Users validate task outcome, usability, exception handling, oversight and real process fit — not merely whether the model produces plausible text.	Business-owner sign-off against measurable outcomes.
Parallel/shadow run	Compare AI recommendations/actions with the existing human/system path for consequential use	Difference/error analysis and calibrated thresholds.

Assurance stage	AI/agentive interpretation	Typical exit evidence
	cases where feasible.	
Canary / phased go-live	Restrict users, transactions, geography, permissions and/or autonomy; progressively expand only when evidence is stable.	Canary metrics and explicit scale decision.
Operational acceptance	Support, CMDB/configuration, SLOs, monitoring, cost budgets, incident/rollback/kill paths, vendor escalation and workforce/legal gates are ready.	OAT checklist and service-owner acceptance.
Continuous production assurance	Regression, drift, tool/model/provider silent update, safety/security monitoring, cost anomalies and periodic adversarial retest.	Continuous evaluation dashboard, provider/version change evidence and reassessment records.

Assurance continues beyond release.



Depth scales with consequence and risk.

The same lifecycle should be **proportionate**: low-risk drafting tools do not require a bank-grade parallel run, while a high-impact autonomous workflow may require all of the above plus independent assurance. [2, 58]

Calibrate evaluation thresholds rather than inventing pass marks

"Evaluation thresholds are met" is not an assurance method unless the thresholds themselves are justified. For each material evaluation, the release record should state **why the threshold is appropriate for the decision and risk**. Depending on the use case, calibration should consider: human or incumbent-system baseline; cost/severity asymmetry between false positives and false negatives; required statistical confidence and sample size; subgroup performance where relevant; inter-annotator agreement for subjective labels; tolerated tail failure; and drift/retest triggers.

A threshold should become more demanding as consequence and irreversibility increase, and it should be revisited when the model, workflow, population or error cost changes. ISO/IEC TS 42119-2 provides a useful risk-based testing frame; the numeric pass mark remains an enterprise decision that needs documented rationale rather than a universal number. [58]

Where an LLM is used as a judge/evaluator, calibrate it against human-labelled examples, report inter-rater agreement where relevant, and measure **multi-run variance** rather than treating one model judgment as ground truth. Keep evaluation-set lineage in the service/model record, protect test sets from prompt/training leakage, and separate RAG retrieval evaluation from answer-generation evaluation so a good final answer cannot hide a weak retrieval layer. [58, 87]

Choose human oversight by risk and operating speed

Human-in-the-loop is not a universal synonym for safe AI. Requiring a person to approve every micro-action can create delay, approval fatigue and false assurance, while some machine-speed failures can complete before manual intervention is practical. Conversely, consequential individual decisions may genuinely require human authorization, review or appeal.

For legally significant automated decisions, oversight is not merely a UX choice. GDPR Article 22 and its safeguards must be assessed now; for high-risk AI, Article 14(4)(b) explicitly requires overseers to remain aware of automation bias and Article 26(2) requires deployers to assign oversight to people with the necessary competence, training and authority. The operating model should therefore provide **information, a meaningful intervention route, contest/appeal, override authority and enough time/context to exercise judgment** where those duties apply. [59, 35, 80, 81, 82]

The blueprint should select the oversight pattern deliberately:

Oversight pattern	Appropriate use
Human-in-the-loop (HITL)	A specific consequential decision/action requires explicit human authorization before execution.
Human-on-the-loop (HOTL)	The system operates within hard limits; humans monitor exceptions, drift, risk signals and policy breaches.
Deterministic automated enforcement	Speed or volume makes manual interception unrealistic; code/policy automatically blocks boundary violations, budget breaches or unsafe states.
Human-out-of-the-loop	Only for sufficiently low-risk, bounded and reversible activity with evidence that the residual risk is acceptable.

Gartner's 2026 agent-governance guidance reinforces the need to distinguish autonomy from access scope rather than governing all agents uniformly. [49]

Match oversight to consequence and speed.



5.8 Operations and service management

Integrate AI with IT service management

One AI-focused project-manager specification in the corpus is particularly valuable here because it treats AI services as production enterprise services rather than innovation artefacts. It calls for solution documentation, security

review, governance approvals, Service Portal offerings, CMDB, business pilot, service-desk handover, support model, runbooks, decision records and clear owners.

That should be standard practice.

Before production, every material AI service should answer:

- Who owns it?
- Who supports it?
- Where is it registered?
- What model/version does it use?
- What systems can it access?
- What data can it see?
- What happens when it fails?
- Who receives an incident?
- How is access revoked?
- How is a model change tested?
- How does it roll back?
- What is its cost?
- When will it be reviewed?

AI should eventually become integrated into normal enterprise architecture, security, change, incident, problem, continuity, vendor and service-management processes—with AI-specific extensions rather than a parallel bureaucracy.

For material production services, extend the CMDB/configuration model so that the service can be reconstructed. Relevant configuration items or linked records typically include:

model/provider/version; embedding model/index build; system prompt or configuration hash; agent/orchestrator version; approved tool/MCP server and version; tool permissions/scopes; policy-set version; evaluation dataset/version; material RAG/data sources; agent/NHI identity; deployment region/end point; and key downstream dependencies.

The **operational acceptance gate** should require evidence that: ownership/support are assigned; risk/security/privacy/workforce gates are closed or formally accepted; evaluation thresholds are met; monitoring and alerting are live; SLOs and execution budgets are configured; access revocation and kill-switch paths have been tested; rollback/recovery includes persistent memory/shared state; runbooks and service-desk escalation are usable; vendor support/escalation exists; continuity/fallback is tested where material; and deletion/access-change propagation has been verified for governed knowledge stores.

For RAG and persistent-memory services, that final evidence should include the **test method, timestamp, affected source/state identifier, entitlement change or deletion event, every derived store checked, observed propagation time, result, and remediation/exception if the SLO was missed**. This makes privacy and memory controls sampleable by operations, risk and Internal Audit rather than merely asserted.

This turns "production-ready" from an architectural opinion into an auditable service-management decision.

Service transition, run and retirement

Production approval should lead into a defined **service-transition and run model**, not end at go-live. For material services, set a hypercare period with explicit exit criteria; name on-call/support ownership proportionate to criticality and operating hours; create problem records for recurring AI failure modes rather than repeatedly closing incidents; and run blameless post-incident reviews that feed corrective actions into the AIMS continual-improvement loop.

Retirement should be as controlled as onboarding. A decommissioning checklist should, as applicable: stop new executions; revoke non-human identities, tokens and scopes; disable tools, webhooks, schedules and integrations; remove or archive endpoints; export/return required data; verify deletion or expiry of RAG indexes, embeddings and

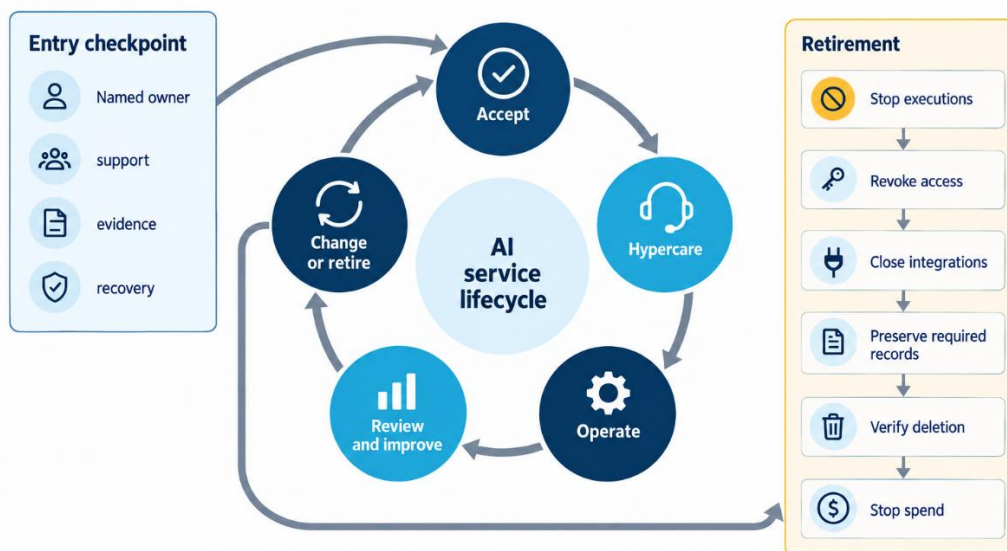
persistent memory; preserve records/legal holds; terminate licences and supplier access; update the AI inventory, CMDB and service catalogue; and verify that no orphaned spend, credentials, queues or dependencies remain.

Records and legal hold: determine which AI outputs, decision traces, prompts/configurations, approvals and agent-generated applications constitute enterprise records for the relevant business process. Retention schedules should support litigation hold, regulatory investigation and defensible disposal without retaining every token forever. Short-lived or employee-created AI applications should still be registered when they influence material decisions or access sensitive data; Gartner's 2026 prediction on intentionally disposable applications specifically highlights records-management and compliance implications. [50]

Accessibility: where an AI-mediated customer service falls within the scope of applicable accessibility law, include accessibility in design and UAT rather than treating the chatbot/agent as exempt because it is AI. In the EU, the European Accessibility Act is **Directive (EU) 2019/882**, applying from 28 June 2025 to specified consumer products and services such as consumer banking and e-commerce. [65]

For third-party or public models, “reproducibility” should normally mean evidence reproducibility rather than a promise of byte-for-byte rerun. Retain the inputs, outputs, model/provider/version, relevant configuration, retrieved-context references, tool calls, timestamps, policy/approval decisions and material human interventions needed to reconstruct the decision path. Where provider retirement makes exact replay impossible, the record should make that limitation explicit.

AI remains an enterprise service after go-live.



5.9 Vendor and third-party management

Treat AI procurement and third-party dependencies as lifecycle controls

AI suppliers can change models, training/data practices, subprocessors, tool chains, geographic processing and pricing after initial approval. Procurement is therefore part of the control system, not merely a sourcing event.

For material model, SaaS, platform and agent/tool suppliers, the contract and vendor record should address:

Treat the **AI gateway, orchestration framework, evaluation stack and agent/tool runtime as Tier-1 software dependencies** where they sit on the path to sensitive data or consequential action. Pin/verify versions, use normal software supply-chain controls, maintain SBOM/ML-BOM/model provenance as applicable, and avoid unreviewed public-package installation in privileged agent environments. The March/April 2026 LiteLLM/axios incidents reinforce that AI infrastructure inherits ordinary package-ecosystem risk. [92, 93, 95]

Contract/control topic	Enterprise expectation

Contract/control topic	Enterprise expectation
Data use and training	State whether enterprise prompts, inputs, outputs, fine-tuning data or telemetry may be retained or used for model/service improvement and under what conditions.
Confidentiality and IP	Confidentiality, ownership/licensing of inputs/outputs/configuration, infringement allocation and restrictions on supplier reuse.
Subprocessors and downstream tools/models	Disclosure/change process for material subprocessors, model providers and agent/tool dependencies.
Security and incidents	Security obligations, vulnerability handling, evidence/cooperation and a contractually defined notification timeframe proportionate to operational/regulatory needs. Do not assume one universal 72-hour clause.
Model/service change	Notice and reassessment triggers for material model, safety, functionality, API, data or policy changes.
Location and transfers	Processing/storage locations, international transfer mechanism and residency constraints where applicable.
Audit/regulatory support	Audit evidence, regulator cooperation, records and right to obtain information proportionate to risk.
Service continuity	Availability/support commitments, minimum deprecation-notice expectation proportionate to criticality , fallback/export mechanisms and critical dependency concentration assessment.
SaaS AI feature change	Require notice/control over material default-on or materially changed AI features that alter data flows, model/provider use, subprocessors, retention, permissions or automated actions.
Portability / switching	Reflect applicable Data Act switching/interoperability rights and Commission model contractual clauses where relevant; test practical export/migration rather than relying on a contractual sentence alone. [78, 94]
Economics	Transparent usage metrics, pricing/change notice and budget protections appropriate to the deal; commercial caps are a negotiated option, not a universal AI control.
Exit and deletion	Data/configuration portability, transition assistance, deletion/return obligations and evidence/attestation where warranted.

For financial entities, overlay DORA ICT third-party requirements and register obligations rather than inventing a separate AI-only supplier regime. [31, 32]

For suppliers of general-purpose AI models, procurement should also record whether the provider relies on/signs the **EU GPAI Code of Practice** and what documentation that makes available to downstream users. The Code is voluntary, but the Commission and AI Board have recognized it as an adequate voluntary mechanism for providers to demonstrate compliance with relevant GPAI obligations; downstream enterprises can therefore use signatory/documentation status as one due-diligence input without treating it as a substitute for their own obligations. [62]

An enterprise that fine-tunes or otherwise modifies a general-purpose AI model should also assess whether the modification changes its regulatory role. The Commission's GPAI guidance uses training-compute thresholds, including a one-third-of-original-training-compute presumption for a significant downstream modification, as part of that analysis. Treat this as a provider-role question, not merely an MLOps change ticket. [111]

Treat sustainability and compute capacity as proportionate architecture/procurement concerns

AI energy and infrastructure demand should be considered where it is material to cost, resilience, procurement commitments, corporate sustainability reporting or stakeholder expectations. The blueprint does **not** impose a universal AI-specific carbon-control regime. Instead, material workloads should capture enough information to support energy/compute-aware model selection, provider due diligence, capacity planning and applicable enterprise sustainability reporting. For many use cases the highest-value action is simply avoiding unnecessarily large models, repeated inference and wasteful agent loops — controls that also improve cost and resilience. Gartner's September

2026 strategic predictions illustrate why energy and AI infrastructure are becoming enterprise-planning concerns rather than purely data-centre issues. [50]

5.10 Economics, funding and AI FinOps

Make AI FinOps a runtime control, not a monthly report

Agentic workloads can produce variable cost because model choice, context length, retry behaviour, tool use and recursive planning can change from one run to the next. Gartner's September 2026 **strategic predictions** recommend linking AI consumption to business value and applying quotas and monitoring controls, while predicting wider adoption of inference-time AI FinOps. The FinOps Foundation likewise treats AI as a workload requiring usage attribution, unit economics and optimization across model, infrastructure and business dimensions. [50, 106]

The runtime loop should be **Attribute → Budget → Alert → Enforce → Optimize**:

Phase	Required control
Attribute	Cost by business unit, use case, workflow, agent, model and material tool/API call.
Budget	Expected cost envelope per run, day and portfolio; separate experimentation from production budgets.
Alert	Detect abnormal token/context growth, retries, tool-call bursts and model-routing anomalies.
Enforce	Deterministic token/cost ceilings, retry and recursion limits, rate limits, model restrictions and circuit breakers.
Optimize	Context pruning/summarization, caching, smaller-model routing, retrieval discipline, batch design and workflow redesign.

The key unit is **cost per successful outcome**, not cost per token in isolation. A cheaper model that causes retries, human remediation or failed transactions may be economically worse than a more expensive model that completes the workflow reliably.

FinOps should also govern **commercial consumption mode**: provisioned/committed capacity versus pay-as-you-go, reserved accelerator capacity, caching commitments, and provider price/model changes. These choices should be reassessed against utilization, resilience, lock-in and workload predictability rather than made once at procurement.

For multi-agent systems, the parent workflow should own the budget envelope. Child agents and delegated sub-workflows should receive bounded sub-budgets for tokens, tool/API spend, retry depth, concurrency and elapsed time; fan-out must stop or degrade gracefully when the remaining parent budget cannot support another branch. This prevents a seemingly cheap planner from spawning an economically uncontrolled execution tree.

Cost controls are also security controls. Gartner's September 2026 research identifies deliberate **cost-exhaustion attacks** in which adversaries drive excessive AI usage to impose operational cost, and recommends treating token consumption and abnormal AI-usage patterns as cybersecurity indicators. The security monitoring stack should therefore alert on anomalous token/context growth, request fan-out, repeated expensive model routing and tool-call bursts even when those events remain within conventional availability thresholds. Hard execution budgets, quotas, rate limits and circuit breakers should feed both FinOps and security operations. [50]



5.11 People, roles and capability building

Roles required in the mature organization

The job-specification corpus points to a notably broader role model than "data scientists plus an AI governance person." The specifications include AI strategists, AI governance leads, AI enablement leads, AI project/program leadership, product roles, platform roles, deployment strategists, agentic architects and AI/ML delivery specialists.

In smaller organisations, one person may double-hat across several role families where competence, independence and segregation-of-duties requirements permit. HR/People and employee-relations capability should remain explicitly available for workforce impact, consultation and role-design decisions rather than being assumed to sit inside generic change management.

A mature capability should therefore contain, or have reliable access to, the following role families:

Role family	Key responsibilities
Chief AI Officer / Head of AI / AI Transformation Director	Enterprise strategy, target state, investment priorities, operating model and executive accountability.
AI Portfolio & Value Lead	Use-case funnel, investment cases, prioritization, benefits, portfolio reporting and stop/scale decisions.
AI Governance & Responsible AI Lead	AI management system, policy, inventory, risk taxonomy, compliance and assurance.
AI Risk / Model Risk specialists	Independent challenge, high-impact model assessment, controls and ongoing risk monitoring.
AI Security Architect / Lead	AI threat models, agent IAM, least privilege, secure integrations, red-team standards and incident readiness.
AI Platform Product Lead	Enterprise model/agent platform as a product, roadmap, user experience and platform economics.
AI / Solution Architects	Reference architectures, integration patterns, RAG, agents, APIs and non-functional requirements.
Agent Product Manager / Agentic Product Lead	Agent-enabled product outcomes, autonomy boundaries, workflow design, user experience, adoption and lifecycle decisions.
AI Evaluation / Quality Engineer	Evaluation design, datasets, thresholds, regression, judge calibration,

Role family	Key responsibilities
	RAG/system testing and release evidence.
AI Red-Team / Adversarial Testing specialist	Threat-led evaluation of prompt injection, data/tool abuse, privilege escalation, exfiltration and multi-agent failure modes.
AI FinOps Analyst / Engineer	Usage attribution, budgets, cost controls, provider economics, optimization and cost-per-outcome analysis.
ML / AI / LLM engineers	Models, GenAI applications, retrieval, evaluation, tools, agents and production code.
MLOps / LLMops / AgentOps engineers	Deployment, versioning, telemetry, evaluation pipelines, release management, rollback and production operations.
Data/Knowledge engineers	Data products, metadata, quality, retrieval corpora, document pipelines and knowledge architecture.
AI Product Owners	Domain discovery, product roadmap, backlog, usability, adoption and outcomes.
Process / Business Analysts	Process decomposition, requirements, baseline metrics, control points and redesigned ways of working.
AI Enablement & Change Lead	Training, communications, champions, role redesign, adoption and behavioral change.
Deployment / Program Leads	Complex cross-functional execution from idea through production and BAU.
AI Service Manager	Catalogue, SLA/SLO, support, ITSM, incidents, change, CMDB/configuration and service lifecycle.
AI Procurement / Vendor Lead	Commercials, licensing, exit provisions, third-party risk and AI-specific supplier assessment.
Legal, Privacy and Compliance specialists	AI Act, GDPR, copyright/IP, contractual obligations and sector-specific regulation.
Internal Audit	Independent assurance over the AI management and control system.

This multidisciplinary requirement is not optional for scaled AI. ISO/IEC 42001 explicitly frames AI management as an organizational management system rather than a narrow technical discipline. [1, 18]

Workforce enablement, training and organizational change

Large-scale AI adoption is principally a **behavior and operating-model change supported by technology**.

The job-specification corpus repeatedly requires the ability to train and enable staff, influence without formal authority, communicate between business and technology, facilitate workshops, build adoption, operate change programmes and translate AI into practical workflows. AI enablement and change therefore belong in the core operating model rather than being appended to deployment.

This is also increasingly a regulatory issue in Europe. The Commission states that the AI Act's AI-literacy duty has applied since **2 February 2025** to providers and deployers. Regulation (EU) 2026/1744 replaced Article 4 while keeping the duty on those operators: it does **not require them to guarantee any specific level of AI literacy of an individual**, while the Commission and Member States support implementation and the AI Board is to issue recommendations. Training should therefore be proportionate to people's knowledge, experience, role and the context in which AI is used. [34, 36]

A single generic two-hour "AI course" is therefore inadequate.

Use a role-based curriculum

Audience	Training required
Board and executive committee	AI economics, strategic opportunities, agentic implications, regulatory obligations, cyber/risk, investment governance and workforce impact.

Audience	Training required
All employees	AI fundamentals, approved tools, prompt practices, verification, confidentiality, privacy, IP, bias, hallucinations, disclosure and escalation.
Managers/process owners	Use-case discovery, process redesign, business cases, human-AI allocation, controls, adoption and benefits realization.
Approvers / human overseers	Decision authority, automation bias, meaningful intervention, contestability, override/reject/escalate duties and the competence/authority required for consequential oversight. [35, 81, 82]
Power users/champions	Advanced prompting, structured workflows, RAG, reusable patterns, evaluation and peer coaching.
Product/project managers	AI lifecycle, requirements, data dependencies, evaluation, uncertainty, risk gates, vendor/platform implications and value tracking.
Engineers/data scientists	Secure AI SDLC, evaluation, RAG, models, agents, tool security, MLOps/LLMOps/AgentOps, monitoring and incident management.
Security/IAM teams	Prompt injection, model/tool attack surfaces, agent identities, permission analysis, supply chain, red teaming and runtime controls.
Legal/risk/privacy/procurement	AI taxonomy, model lifecycle, AI Act, GDPR, contracts, technical controls, vendor models and assurance evidence.
Service desk/operations	AI services, known failure modes, incidents, escalation, access, rollback and end-user support.
Internal Audit	AI management system, lifecycle evidence, controls, model/agent risks and assurance techniques.

The correct competence is therefore not "knows prompt engineering." It is:

knows when AI is useful, can structure the interaction, can assess whether the result is trustworthy enough for the purpose, and understands what must remain a human responsibility.



Train people around workflows rather than demonstrations

Generic examples generate enthusiasm; workflow-specific training changes performance.

For each scaled use case, employees should receive:

- the old and new workflow;
- which work AI performs;
- which work remains theirs;
- verification expectations;

- escalation thresholds;
- prohibited uses;
- examples of good and bad outputs;
- expected productivity or quality target;
- how user feedback improves the service.

Managers need additional training on capacity realization. Otherwise a company can create thousands of small time savings while changing neither headcount, throughput nor customer outcomes.

Establish an AI champions network

Every significant function should have trained local champions who:

translate central policy into local context; discover use cases; coach colleagues; gather feedback; flag shadow AI; share reusable patterns; and identify skills gaps.

Champions should not become amateur compliance officers. They are an adoption and communication layer between the central function and business domain.

5.12 Workforce, employment and organisational change

Treat employment relations and workforce governance as part of deployment

Training and adoption are only part of the workforce problem. AI can change duties, reporting lines, work allocation, staffing assumptions, performance measurement and the degree to which employees are monitored. Those changes may create legal and collective-relations obligations that must be integrated into delivery governance rather than left to post-launch change management.

The enterprise should add a **workforce impact assessment** for material deployments covering: roles/tasks affected; decisions delegated to AI; monitoring/telemetry involving employees; skills and redeployment; expected capacity release; changes to working conditions; equality/fairness impacts; communication; consultation/negotiation obligations; and accountable HR/legal/employee-relations owners. Agent traces, productivity telemetry and review logs that identify employee behaviour should be screened as **employee monitoring**, with purpose, proportionality, access/use, retention and privacy/consultation implications assessed rather than treating telemetry as operationally neutral.

In Sweden, MBL 11 § is a concrete example: for employers bound by a collective agreement, important changes to the business or important changes to working/employment conditions trigger an employer-initiated negotiation obligation before the decision. [38] The blueprint therefore uses a **consultation applicability gate**, not the overbroad claim that every AI deployment automatically requires MBL negotiation.

Where employee-related processing is likely to create high risk to individuals' rights and freedoms, privacy assessment — including a DPIA where required — should run alongside the employment-relations process rather than after technical deployment. The AI Act does not displace GDPR responsibilities. [39]

A practical **HR/legal deployment path** is:

1. **Screen the workforce impact:** roles, task allocation, employment decisions, monitoring, productivity scoring, surveillance, staffing or working-condition changes.
2. **Classify legal/regulatory status:** determine high-risk status, operator role and application date; separately test whether Article 27 FRIA applies. For an ordinary private employer, Annex III employment classification alone does not automatically create an FRIA duty. [35]
3. **Assess information/consultation duties:** apply the jurisdiction-specific consultation rules identified by the workforce gate. In Sweden, use the MBL analysis above; the trigger is the significance of the proposed change, not the fact that the technology is AI. [38]
4. **Assess privacy and fundamental rights:** legal basis, necessity/proportionality, transparency, employee-data rules and a DPIA where the processing is likely to create high risk; complete a distinct FRIA only where

Article 27 scope applies. Member-State employment rules under GDPR Article 88 may also be relevant. [35, 39, 59]

5. **Prepare worker information where applicable:** Article 26(7) will require information to workers' representatives and affected workers before use of high-risk workplace AI once the relevant high-risk provisions apply. [35, 36]
6. **Retain evidence:** impact assessment, consultation/negotiation record where applicable, privacy decision, communications, training and final accountable approval.

Do not use Swedish **MBL §38** as a generic AI-monitoring trigger: that provision concerns certain decisions to have work performed by people who are not employees of the employer. [38]

More generally, national employment, privacy and collective-agreement rules can create information, consultation, assessment or negotiation duties **earlier than the AI Act's later high-risk-system dates**. The deployment gate should therefore evaluate the law applicable to the actual workforce and processing context rather than using the AI Act timetable as the only trigger.

Redesign jobs, not just tasks

As AI becomes pervasive, workforce planning should explicitly examine:

tasks removed; tasks augmented; tasks newly created; decisions that remain human; supervisory responsibilities over agents; skills that become more valuable; capacity released; organizational layers/handoffs that can disappear.

Recent research on AI-enabled product-development organizations suggests that embedding tools alone produces much less differentiation than combining AI with changes to roles, workflows and decision rights. [22]

Job redesign should also protect the **future skills pipeline**. If AI absorbs entry-level drafting, analysis or coding tasks, deliberately preserve supervised practice, rotations, review duties and opportunities to build judgment rather than optimizing away the experiences from which future experts are formed. Emerging labour-market evidence on more AI-exposed early-career roles makes this a workforce-planning risk worth monitoring, not a settled forecast of net employment effects. [107]

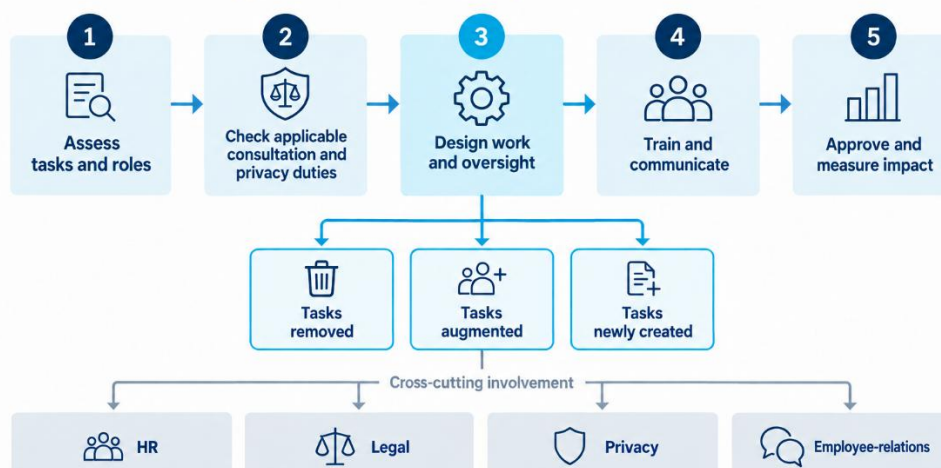
The ultimate workforce question is therefore not:

"How many jobs will AI automate?"

but:

"What is the optimal allocation of sensing, retrieval, generation, calculation, prediction, judgment, approval, relationship management and execution between humans, traditional software, ML and agents?"

Redesign work with people and evidence.



PART VI — CONCLUSION

Enterprise AI maturity is not the number of tools deployed, users enabled or agents created. It is the organisation's ability to select the right problems, fund them, deliver them repeatedly, constrain them safely, operate them through normal enterprise processes and demonstrate realised value.

The most scalable model is therefore additive rather than revolutionary: extend the organisation's existing strategy, governance, security, data, model-risk, engineering, procurement, service-management, finance and workforce systems; create dedicated AI capabilities only where genuinely new control objects or shared services require them; and increase delegated autonomy only as evidence and domain capability mature.

The end state is an AI-pervasive enterprise in which AI is a normal design option, not an exception and not an ideology. The operating model should make useful AI easier to deploy, unsafe AI harder to deploy, low-value AI easier to stop and the evidence behind all four decisions visible.

PART VII — APPENDICES AND PRACTITIONER TOOLKIT

Appendix A — Lift-and-adapt practitioner artefacts

Lift-and-adapt practitioner artefacts

The blueprint is intended to be implemented, not merely referenced. The following compact artefacts consolidate controls already described throughout the document into forms that a programme team can lift and tailor. They are **starting templates**, not substitutes for sector- or jurisdiction-specific requirements.

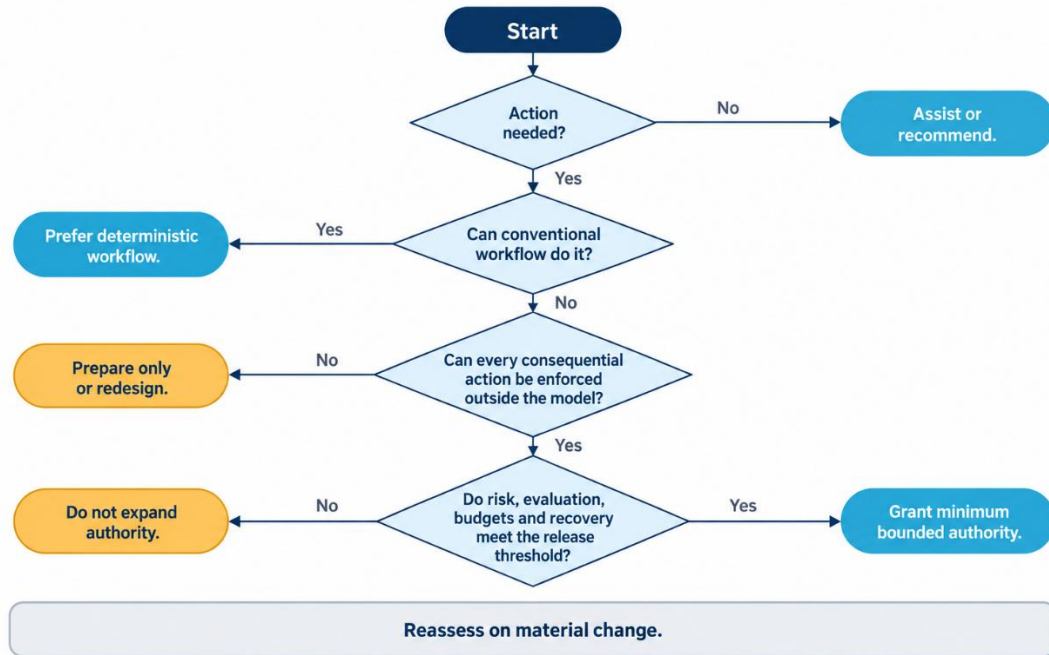
One-page AI opportunity card

Field	Minimum content
Owner and problem	Business/process owner; workflow/problem; affected users/customers.
Baseline and target	Current volume, cycle time, quality/cost/risk baseline; measurable target outcome.
Solution pattern	Deterministic automation, predictive ML, GenAI/RAG, AI-assisted workflow or agentic design; explain why simpler alternatives are insufficient.
Data and authority	Data classes/sources; integrations/tools; read/write/action authority; human responsibility.
Risk and legal	Internal risk/autonomy tier; AI Act operator role (where applicable); applicable AI/privacy/security/workforce/sector obligations; FRIA/DPIA links where required.
Economics	Full lifecycle TCO; benefit range; expected cost per successful outcome.
Capacity/value conversion	Specific mechanism by which time/capacity becomes throughput, avoided hiring, redeployment, SLA improvement, revenue or cost reduction; named owner.
Measurement and exit	Pilot comparison design; thresholds; scale/stop criteria; rollback/retirement route.

Agent-autonomy decision tree

1. **Does the use case require the system to take an action?** If no, remain Assist/Recommend.
2. **Can the action be implemented deterministically without agent planning?** If yes, prefer conventional workflow/rules.
3. **Is the action reversible and low consequence?** If no, require explicit approval or a stronger assurance case.
4. **Can identity, purpose, scope, transaction limits and segregation of duties be enforced outside the model, and does every consequential action traverse a policy-enforced API or equivalent deterministic boundary?** If no, do not grant action authority. For UI-acting agents, cap authority at **Prepare** unless an equivalent isolated/enforced boundary is demonstrably in place.
5. **Have evaluation, adversarial/system testing, cost budgets, monitoring, rollback/kill and service ownership met the applicable release threshold?** If no, do not increase autonomy.
6. **Expand authority only to the minimum level needed for the measured business outcome**, and re-assess on material change.

Should this agent receive action authority?



Operational-acceptance checklist

Before production, confirm at minimum: named service/business owner and support/on-call route; inventory/CMDB record; approved model/tool/configuration versions; closed or accepted risk/security/privacy/workforce gates; justified evaluation thresholds and negative-authorization results; SLOs/alerts; execution budgets; access-revocation and kill tests; rollback including memory/shared state; deletion/access-change propagation where applicable; vendor escalation; continuity/fallback; incident/runbook and notification-clock readiness; records/retention/legal-hold classification; accessibility testing where legally/applicably in scope; **defined hypercare and exit criteria**; and defined review/retirement owner/date.

Decision-rights matrix — minimum fields

Lift the organizational-model matrix into the local operating model and map enterprise titles to it. At minimum record: **decision; accountable authority; required consult/review roles; evidence required; delegation limit; escalation route; exception authority; and retained decision record**. Do not combine approval and independent assurance where the selected risk tier requires separation of duties.

Decommissioning checklist

For retirement, record at minimum: stop-new-execution decision and owner; disabled schedules/webhooks/endpoints; revoked NHI identities/tokens/scopes; supplier/licence closure; exported/returned data; verified deletion/expiry of RAG indexes, embeddings and memory where applicable; archived records/legal holds; updated CMDB/service catalogue/AI inventory; cleared queues/orphaned credentials; residual dependencies; final cost stop; and decommission evidence/approval.

Material-change trigger list

Reassessment is triggered by a material change in **purpose; affected population; model/provider/version; system prompt/agent logic; training/evaluation data; RAG corpus; memory design; tool/server/version; API scope; read/write privilege; delegated authority; transaction limit; geography/residency; external exposure; downstream integration; human oversight; supplier/subprocessor; or autonomy level**. The original approval applies to a defined system configuration and use, not to unlimited future evolution.

AIMS evidence-pack skeleton

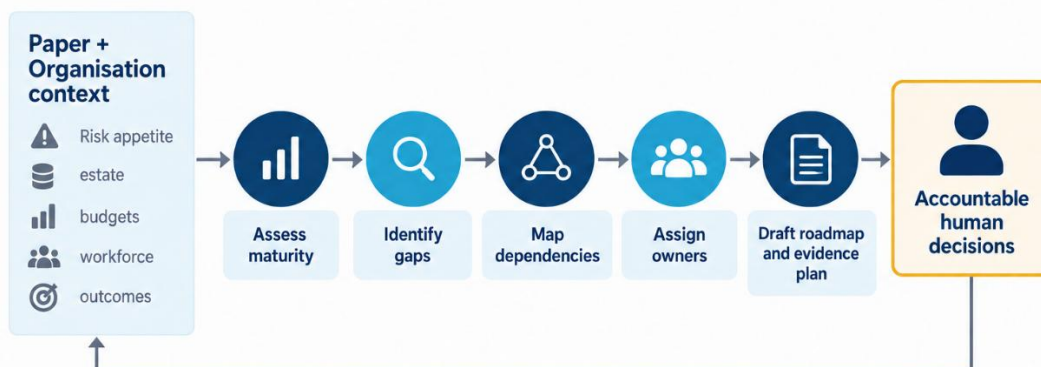
Maintain one evidence index linking: AIMS scope/context and applicable obligations; policy/objectives/decision rights; AI inventory; risk and impact assessments; competence/literacy evidence; architecture/security/privacy/workforce decisions; supplier due diligence; versioned evaluations and red-team results; service/OAT records; performance/cost/value metrics; incidents/nonconformities; internal-audit results; management-review decisions; corrective actions; exceptions/waivers; and retirement evidence. The index should point to authoritative records rather than duplicate every artefact into another repository.

Appendix B — Machine-assisted use of this paper

A human or authorised governance/planning agent can use this paper as an umbrella reference by combining it with organisation-specific context: current operating model, regulatory perimeter, risk appetite, technology estate, AI inventory, budgets, workforce constraints and target outcomes.

A useful machine-assisted workflow is: (1) assess the organisation against the Part IV maturity dimensions; (2) identify capability gaps and dependencies; (3) map them onto the Part II journey; (4) assign accountable enterprise owners using existing structures wherever possible; (5) produce a sequenced roadmap, evidence plan and decision log; and (6) route legal, regulatory, security, employment and risk judgements to accountable humans. The agent should cite the relevant paper section and source rather than treating the paper as authority in its own right.

Turn the paper into an implementation plan.



Cite paper sections and sources; the agent does not determine compliance.

Appendix C — Research method and assurance boundary

Review methodology and residual assurance boundary

This research paper has been strengthened through deep research plus several independent AI-assisted adversarial reviews. Each review was treated as **input, not authority**: material legal or time-sensitive claims were checked against primary/authoritative sources before integration, and several reviewer assertions were rejected or narrowed when they did not match the source law.

That iterative process improves breadth and internal challenge. The document should not be represented as legal advice, regulatory approval or independent audit. Before an organization relies on it for a regulated deployment, obtain human SME validation appropriate to the context — for example employment counsel, privacy/legal, model-risk, cybersecurity, sector regulation and IT service/operational assurance.

Appendix D — Implementation anti-patterns

Implementation anti-patterns to avoid

Anti-pattern	Why it fails	Better pattern
Agent self-policing — "the prompt tells it not to do X"	Probabilistic instructions are not a reliable authorization boundary.	Deterministic IAM/policy enforcement at the action boundary.
Autonomy by default	Increases blast radius before value/reliability are demonstrated.	No action authority by default; expand only with evidence.
Uniform governance for every AI use case	Creates bottlenecks for low-risk use while under-specifying genuinely critical systems.	Risk- and autonomy-tiered lifecycle.
Time saved = cash benefit	Small time savings often do not translate into realized capacity or P&L.	Named benefit owner and explicit capacity/value conversion mechanism.
Human rubber stamp	Approval fatigue and automation bias can make nominal HITL weaker than hard controls.	Reserve human approval for consequential judgment; automate enforceable boundaries.
Pilot success = production readiness	A demo can work without support, security, rollback, economics or adoption.	Separate technical pilot, business acceptance and operational acceptance gates.
Static approval of changing models/tools	Model, tool, scope or permission drift invalidates the original assurance case.	Material-change triggers, versioned configuration and reassessment.
AI exception to enterprise operations	Creates unowned services and unauditable dependencies.	Extend normal architecture, cyber, procurement and ITSM with AI-specific controls.
More agents = more value	Multi-agent topology increases cost, latency and failure propagation.	Use the simplest architecture capable of delivering the required outcome.
Developer convenience credentials	Coding/infra agents inherit broad shell, cloud or production credentials, collapsing separation between probabilistic planning and privileged execution.	Short-lived scoped identities, sandboxing, protected branches and independent deployment gates.
Agent in a human session	UI/browser agents inherit the human's full authenticated authority and can bypass the policy plane.	Separate agent identity/session, isolated browser/desktop, allowlists/recording; cap at Prepare where authority cannot be deterministically bounded.

Appendix E — Related Cognostrix research

Agentic AI Controls, Governance and Assurance — A Reference Catalogue, v1.0 (September 2026): detailed catalogue of agent control mechanisms, implementation patterns, testing and assurance evidence.

<https://cognostrix.com/news/agentic-ai-control-catalogue-v1.html>

<https://cognostrix.com/docs/Cognostrix-Agentic-AI-Control-Catalogue-v1.0.pdf>

References and source trail

Public sources

The reference list is cumulative across versions. Primary law/regulator/standards sources are preferred for requirements; research and practitioner evidence is retained where it materially supports implementation or emerging-risk analysis.

1. ISO/IEC 42001:2023 - Artificial intelligence — Management system. <https://www.iso.org/standard/81230.html>
2. Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile | NIST. <https://www.nist.gov/publications/artificial-intelligence-risk-management-framework-generative-artificial-intelligence>
3. What Is an AI Center of Excellence? | IBM. <https://www.ibm.com/think/topics/ai-center-of-excellence>
4. Seizing the agentic AI advantage | McKinsey. <https://www.mckinsey.com/capabilities/quantumblack/our-insights/seizing-the-agentic-ai-advantage>
5. OWASP Top 10 for Agentic Applications for 2026 - OWASP Gen AI Security Project. <https://genai.owasp.org/resource/owasp-top-10-for-agentic-applications-for-2026/>
6. AI Act | Shaping Europe's digital future. <https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai>
7. The enforcement framework of the AI Act | Shaping Europe's digital future. <https://digital-strategy.ec.europa.eu/en/policies/enforcement-ai-act>
8. Transparency obligations under Article 50 of the AI Act | Shaping Europe's digital future. <https://digital-strategy.ec.europa.eu/en/faqs/transparency-obligations-under-article-50-ai-act>
9. NIST AI RMF Playbook | NIST. <https://www.nist.gov/itl/ai-risk-management-framework/nist-ai-rmf-playbook>
10. Generative AI at Work | NBER. <https://www.nber.org/papers/w31161>
11. Navigating the Jagged Technological Frontier | Harvard Business School AI Institute. <https://aiinstitute.hbs.edu/navigating-the-jagged-technological-frontier/>
12. Shifting Work Patterns with Generative AI | NBER. <https://www.nber.org/papers/w33795>
13. AI Risk Management Framework | NIST. <https://www.nist.gov/itl/ai-risk-management-framework>
14. Least privilege for AI agents with Microsoft Entra Agent ID | Microsoft Learn. <https://learn.microsoft.com/en-us/security/zero-trust/sfi/least-privilege-for-ai-agents>
15. Secure autonomous agentic AI systems | Microsoft Learn. <https://learn.microsoft.com/en-us/security/zero-trust/sfi/secure-agentic-systems>
16. Commission starts enforcing AI Act rules and new transparency requirements on 2 August | Shaping Europe's digital future. <https://digital-strategy.ec.europa.eu/en/news/commission-starts-enforcing-ai-act-rules-and-new-transparency-requirements-2-august>
17. AI at work but not at scale. <https://www.mckinsey.com/featured-insights/charts/ai-at-work-but-not-at-scale>
18. ISO - AI management systems: What businesses need to know. <https://www.iso.org/artificial-intelligence/ai-management-systems>
19. Scaling agentic AI with data transformations | McKinsey. <https://www.mckinsey.com/capabilities/mckinsey-technology/our-insights/building-the-foundations-for-agentic-ai-at-scale>
20. Reimagining tech infrastructure for agentic AI | McKinsey. <https://www.mckinsey.com/capabilities/mckinsey-technology/our-insights/reimagining-tech-infrastructure-for-and-with-agentic-ai>
21. The symbiotic enterprise: A new model for growth | McKinsey. <https://www.mckinsey.com/fr/our-insights/the-symbiotic-enterprise>

22. Rethinking agentic product development | McKinsey. <https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/beyond-the-copilot-scaling-the-agentic-product-development-life-cycle>
23. AIUC-1: Crosswalks OWASP Top 10 For Agentic Applications - OWASP Gen AI Security Project. <https://genai.owasp.org/resource/aiuc-1-crosswalks-owasp-top-10-for-agentic-applications/>
24. OWASP GenAI Security Project Unveils 2026 Top 10 for LLM Applications, New Agent Control Standard and Sponsors as Community Tops 30,000 Members - OWASP Gen AI Security Project. <https://genai.owasp.org/2026/09/01/owasp-genai-security-project-unveils-2026-top-10-for-llm-applications-new-agent-control-standard-and-sponsors-as-community-tops-30000-members/>
25. OWASP GenAI LLM Top 10 2026 - OWASP Gen AI Security Project. <https://genai.owasp.org/resource/owasp-genai-llm-top-10-2026/>
26. Agentic AI | NIST. <https://www.nist.gov/agentic-ai>
27. New Concept Paper on Identity and Authority of Software Agents | NIST. <https://www.nist.gov/news-events/news/2026/02/new-concept-paper-identity-and-authority-software-agents>
28. Announcing the "AI Agent Standards Initiative" for Interoperable and Secure Innovation | NIST. <https://www.nist.gov/news-events/news/2026/02/announcing-ai-agent-standards-initiative-interoperable-and-secure>
29. EDPB opinion on AI models: GDPR principles support responsible AI | European Data Protection Board. https://www.edpb.europa.eu/news/edpb-opinion-on-ai-models-gdpr-principles-support-responsible-ai_en
30. Enhancing compliance and consistency: EDPB adopts DPIA template | European Data Protection Board. https://www.edpb.europa.eu/news/enhancing-compliance-and-consistency-edpb-adopts-dpia-template_ga
31. Preparations for reporting of DORA registers of information | European Banking Authority. <https://eba.europa.eu/activities/direct-supervision-and-oversight/digital-operational-resilience-act/preparation-dora-application>
32. The EBA amends its Guidelines on ICT and security risk management measures in the context of DORA application | European Banking Authority. <https://www.eba.europa.eu/publications-and-media/press-releases/eba-amends-its-guidelines-ict-and-security-risk-management-measures-context-dora-application>
33. Secure AI Framework (SAIF) | Google Cloud. <https://cloud.google.com/use-cases/secure-ai-framework>
34. AI talent, skills and literacy | Shaping Europe's digital future. <https://digital-strategy.ec.europa.eu/en/policies/ai-talent-skills-and-literacy>
35. Regulation (EU) 2024/1689 - consolidated version 27 July 2026. <https://eur-lex.europa.eu/eli/reg/2024/1689/2026-07-27/eng>
36. Regulation (EU) 2026/1744 - Digital Omnibus on AI. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ%3AL_202601744
37. Uppdrag att vara nationella behöriga myndigheter enligt AI-förordningen. <https://www.regeringen.se/regeringsuppdrag/2026/06/uppdrag-att-vara-nationella-behoriga-myndigheter-enligt-ai-forordningen/>
38. Lag (1976:580) om medbestämmande i arbetslivet. https://www.riksdagen.se/sv/dokument-och-lagar/dokument/svensk-forfattningssamling/lag-1976580-om-medbestammande-i-arbetslivet_sfs-1976-580/
39. Opinion 28/2024 on data protection aspects related to AI models. https://www.edpb.europa.eu/documents/opinion-of-the-board-art-64/opinion-282024-on-certain-data-protection-aspects-related-to_en
40. Is My Data in Your Retrieval Database? Membership Inference Attacks Against Retrieval Augmented Generation. <https://arxiv.org/abs/2405.20446>
41. Accelerating the Adoption of Software and AI Agent Identity and Authorization - Concept Paper. <https://csrc.nist.gov/pubs/other/2026/02/05/accelerating-the-adoption-of-software-and-ai-agent/ipd>

42. MCPTox: A Benchmark for Tool Poisoning on Real-World MCP Servers.
<https://ojs.aaai.org/index.php/AAAI/article/view/40895>
43. Model Context Protocol - Transports specification (2025-06-18) — historical version retained for provenance; superseded by the 2026-07-28 specification.
<https://modelcontextprotocol.io/specification/2025-06-18/basic/transports>
44. MCP Supply Chain Advisory: RCE Vulnerabilities Across the AI Ecosystem.
<https://www.ox.security/blog/mcp-supply-chain-advisory-rce-vulnerabilities-across-the-ai-ecosystem/>
45. OMNI-LEAK: Orchestrator Multi-Agent Network Induced Data Leakage. <https://arxiv.org/abs/2602.13477>
46. Open Challenges in Multi-Agent Security: Towards Secure Systems of Interacting AI Agents.
<https://arxiv.org/abs/2505.02077>
47. AgentCheck: A Reproduce-Intervene-Mitigate Workbench for LLM Agents over MCP.
<https://arxiv.org/abs/2607.11098>
48. Gartner Predicts Over 40% of Agentic AI Projects Will Be Canceled by End of 2027.
<https://www.gartner.com/en/newsroom/press-releases/2025-06-25-gartner-predicts-over-40-percent-of-agentic-ai-projects-will-be-canceled-by-end-of-2027>
49. Gartner Says Applying Uniform Governance Across AI Agents Will Lead to Enterprise AI Agent Failure.
<https://www.gartner.com/en/newsroom/press-releases/2026-05-26-gartner-says-applying-uniform-governance-across-ai-agents-will-lead-to-enterprise-ai-agent-failure>
50. Gartner Unveils Top Strategic Predictions for 2027 and Beyond.
<https://www.gartner.com/en/newsroom/press-releases/2026-09-15-gartner-unveils-top-strategic-predictions-for-2027-and-beyond>
51. C2PA Technical Specification 2.4 - Content Credentials.
https://spec.c2pa.org/specifications/specifications/2.4/specs/C2PA_Specification.html
52. The 2026-07-28 Model Context Protocol Specification. <https://blog.modelcontextprotocol.io/posts/2026-07-28/>
53. Enterprise-Managed Authorization: Zero-touch OAuth for MCP.
<https://blog.modelcontextprotocol.io/posts/enterprise-managed-auth/>
54. Secure Production Identity Framework for Everyone (SPIFFE) standard. <https://spiffe.io/docs/latest/spiffe-specs/spiffe/>
55. RFC 8693: OAuth 2.0 Token Exchange. <https://www.rfc-editor.org/rfc/rfc8693.html>
56. ISO/IEC 23894:2023 — Artificial intelligence — Guidance on risk management.
<https://www.iso.org/standard/77304.html>
57. ISO/IEC 42005:2025 — Artificial intelligence — AI system impact assessment.
<https://www.iso.org/standard/44545.html>
58. ISO/IEC TS 42119-2:2025 — Artificial intelligence — Testing of AI — Part 2: Overview of testing AI systems.
<https://www.iso.org/standard/84127.html>
59. Regulation (EU) 2016/679 (General Data Protection Regulation). <https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng>
60. Directive (EU) 2022/2555 (NIS2 Directive). <https://eur-lex.europa.eu/eli/dir/2022/2555/oj/eng>
61. Agent Control Standard (ACS) | OWASP Gen AI Security Project. <https://genai.owasp.org/resource/agent-control-standard-acs/>
62. The General-Purpose AI Code of Practice | European Commission. <https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai>
63. The state of AI in 2026: On the road to ROI | McKinsey.
<https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai>
64. Gartner Predicts by 2027, 50% of Enterprises Without a People-Centric AI Strategy Will Lose Their Top AI Talent. <https://www.gartner.com/en/newsroom/press-releases/2026-05-13-gartner-predicts-by-2027-50-percent-of-enterprises-without-a-people-centric-ai-strategy-will-lose-their-top-ai-talent>

65. Directive (EU) 2019/882 — accessibility requirements for products and services (European Accessibility Act). <https://eur-lex.europa.eu/eli/dir/2019/882>
66. Agentic MCP Security Best Practices Guide | Cloud Security Alliance. <https://labs.cloudsecurityalliance.org/agentic/agentic-mcp-security-best-practices-v1/>
67. The IIA's Three Lines Model | The Institute of Internal Auditors. <https://www.theiia.org/en/content/position-papers/2020/the-iias-three-lines-model-an-update-of-the-three-lines-of-defense/>
68. SS1/23 — Model risk management principles for banks | Bank of England / Prudential Regulation Authority. <https://www.bankofengland.co.uk/prudential-regulation/publication/2023/may/model-risk-management-principles-for-banks-ss>
69. Brynjolfsson, Li and Raymond, "Generative AI at Work," *Quarterly Journal of Economics* 140(2), 2025, pp. 889–942. <https://doi.org/10.1093/qje/qjae044>
70. METR, Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity, 10 July 2025. https://metr.org/Early_2025_AI_Experienced_OS_Devs_Study-paper.pdf
71. METR, We Are Changing Our Developer Productivity Experiment Design, 24 February 2026. <https://metr.org/blog/2026-02-24-uplift-update/>
72. Humlum and Vestergaard, "Still Waters, Rapid Currents: The Hidden Transformation of the Labor Market through Generative AI," NBER Working Paper 33777, revised March 2026. <https://www.nber.org/papers/w33777>
73. Niederhoffer et al., "AI-Generated 'Workshop' Is Destroying Productivity," *Harvard Business Review*, 22 September 2025. <https://hbr.org/2025/09/ai-generated-workshop-is-destroying-productivity>
74. European Commission, Code of Practice on Transparency of AI-generated Content, final 10 June 2026. <https://digital-strategy.ec.europa.eu/en/policies/code-practice-ai-generated-content>
75. European Commission, Guidelines on transparency obligations for providers and deployers of AI systems under Article 50 of the AI Act, 20 July 2026. <https://digital-strategy.ec.europa.eu/en/library/guidelines-transparency-obligations-providers-and-deployers-ai-systems>
76. European Commission, Cyber Resilience Act reporting obligations. <https://digital-strategy.ec.europa.eu/en/policies/cra-reporting>
77. Directive (EU) 2024/2853 on liability for defective products. <https://eur-lex.europa.eu/eli/dir/2024/2853/oj/eng>
78. Regulation (EU) 2023/2854 (Data Act). <https://eur-lex.europa.eu/eli/reg/2023/2854/oj/eng>
79. European Commission, Withdrawal of Commission proposals, OJ C/2025/5423, 6 October 2025. <https://eur-lex.europa.eu/eli/C/2025/5423/oj/eng>
80. Court of Justice of the European Union, C-634/21 SCHUFA Holding (Scoring), judgment/press release 186/23, 7 December 2023. <https://curia.europa.eu/jcms/upload/docs/application/pdf/2023-12/cp230186en.pdf>
81. European Data Protection Supervisor, TechDispatch #2/2025: Human Oversight of Automated Decision-Making, 23 September 2025, with 2026 checklist material. https://www.edps.europa.eu/data-protection/our-work/publications/techdispatch/2025-09-23-techdispatch-22025-human-oversight-automated-making_en
82. Dutch Data Protection Authority, Tools for meaningful human intervention in algorithmic decision-making, 23 July 2025. <https://autoriteitpersoonsgegevens.nl/en/documents/tools-meaningful-human-intervention>
83. Post- och telestyrelsen (PTS), AI-förordningen — Swedish competent-authority information. <https://www.pts.se/ai/ai-forordningen/>
84. Regeringskansliet, Remiss av SOU 2025:101 Anpassningar till AI-förordningen. <https://www.regeringen.se/remisser/2025/11/remiss-av-sou-2025101-anpassningar-till-ai-forordningen/>
85. Post- och telestyrelsen (PTS), Regulatorisk sandlåda för AI. <https://www.pts.se/ai/ai-forordningen/regulatorisk-sandlada-for-ai/>

86. Financial Reporting Council, UK Corporate Governance Code 2024.
<https://www.frc.org.uk/library/standards-codes-policy/corporate-governance/uk-corporate-governance-code/>
87. NIST AI 100-2 E2025, *Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations*, March 2025. <https://csrc.nist.gov/pubs/ai/100/2/e2025/final>
88. MITRE ATLAS — Adversarial Threat Landscape for Artificial-Intelligence Systems. <https://atlas.mitre.org/>
89. NCSC/CISA et al., Guidelines for Secure AI System Development, 27 November 2023.
<https://www.ncsc.gov.uk/collection/guidelines-secure-ai-system-development>
90. ETSI TS 104 223 V1.1.1 (2025-04), Securing Artificial Intelligence (SAI); Baseline Cyber Security Requirements for AI Models and Systems.
https://www.etsi.org/deliver/etsi_TS/104200_104299/104223/01.01.01_60/ts_104223v010101p.pdf
91. NSA/CISA and partners, AI Data Security: Best Practices for Securing Data Used to Train & Operate AI Systems, 22 May 2025. https://media.defense.gov/2025/May/22/2003720601/-1/-1/0/CSI_AI_DATA_SECURITY.PDF
92. LiteLLM, Security Update — March 2026 supply-chain compromise.
<https://github.com/BerriAI/litellm/issues/24518>
93. Microsoft Security Blog, Mitigating the Axios npm supply chain compromise, 1 April 2026.
<https://www.microsoft.com/en-us/security/blog/2026/04/01/mitigating-the-axios-npm-supply-chain-compromise/>
94. European Commission, Updated EU AI Model Contractual Clauses (MCC-AI), 5 March 2025. <https://public-buyers-community.ec.europa.eu/communities/procurement-ai/resources/updated-eu-ai-model-contractual-clauses>
95. Spracklen et al., "We Have a Package for You! A Comprehensive Analysis of Package Hallucinations by Code Generating LLMs," USENIX Security 2025.
<https://www.usenix.org/conference/usenixsecurity25/presentation/spracklen>
96. CycloneDX, Machine Learning Bill of Materials (AI/ML-BOM).
<https://www.cyclonedx.org/capabilities/mlbom/>
97. Amazon Web Services, Security Bulletin AWS-2025-015, 23 July 2025.
<https://aws.amazon.com/security/security-bulletins/AWS-2025-015/>
98. Wiz, s1ngularity's aftermath: Nx supply-chain attack analysis, 27 August 2025.
<https://www.wiz.io/blog/s1ngularitys-aftermath>
99. The Register, Replit AI agent deleted a production database despite code freeze, 21 July 2025.
https://www.theregister.com/2025/07/21/replit_saastr_vibe_coding_incident/
100. The Register, Cursor/Opus agent deleted PocketOS production database and backups, 27 April 2026.
https://www.theregister.com/2026/04/27/cursoropus_agent_snuffs_out_pocketos/
101. European Data Protection Board, Guidelines 03/2026 on web scraping in the context of generative AI (draft; open for feedback at the document date). https://www.edpb.europa.eu/public-consultations/guidelines-032026-on-web-scraping-in-the-context-of-generative-ai_en
102. European Commission, Explanatory Notice and Template for the Public Summary of Training Content for general-purpose AI models, 24 July 2025. <https://digital-strategy.ec.europa.eu/en/library/explanatory-notice-and-template-public-summary-training-content-general-purpose-ai-models>
103. OpenTelemetry, Generative AI semantic conventions (Development status).
<https://opentelemetry.io/docs/specs/semconv/registry/attributes/gen-ai/>
104. OECD, AI Incidents and Hazards Monitor. <https://oecd.ai/en/incidents>
105. AI Incident Database. <https://incidentdatabase.ai/>
106. FinOps Foundation, FinOps for AI Overview. <https://www.finops.org/wg/finops-for-ai-overview/>

107. Stanford Digital Economy Lab, "Canaries in the Coal Mine? Six Facts about the Recent Employment Effects of Artificial Intelligence," revised 12 August 2026. <https://digitaleconomy.stanford.edu/publication/canaries-in-the-coal-mine-six-facts-about-the-recent-employment-effects-of-artificial-intelligence/>
108. European Commission, Draft Guidelines on classification of high-risk AI systems, 19 May 2026. <https://digital-strategy.ec.europa.eu/en/library/draft-commission-guidelines-classification-high-risk-ai-systems>
109. European Commission, Standardisation of the AI Act. <https://digital-strategy.ec.europa.eu/en/policies/ai-act-standardisation>
110. CEN-CENELEC, EN 18286 in the Spotlight: Supporting Compliance with the AI Act, 31 July 2026. <https://www.cencenelec.eu/news-events/news/2026/en-in-the-spotlight/2026-07-30-ai-quality-management/>
111. European Commission, Guidelines on the scope of obligations for providers of general-purpose AI models under the AI Act, 18 July 2025. <https://digital-strategy.ec.europa.eu/en/library/guidelines-scope-obligations-providers-general-purpose-ai-models-under-ai-act>
112. OX Security, The Mother of All AI Supply Chains: Critical Systemic Vulnerability at the Core of the MCP, 15 April 2026. <https://www.ox.security/blog/the-mother-of-all-ai-supply-chains-critical-systemic-vulnerability-at-the-core-of-the-mcp/>

Evidence interpretation note

- **Legal requirement:** legislation, consolidated legal text or regulator/authority guidance.
- **Formal standard/guidance:** standards bodies and recognized frameworks; exact certification mapping should use licensed normative text where applicable.
- **Established practice:** mature vendor/platform or widely adopted engineering guidance; implementation examples are not automatically universal requirements.
- **Emerging evidence:** peer-reviewed/preprint research and current security findings used to identify risks and candidate controls; not automatically treated as universal requirements.
- **Blueprint synthesis:** recommendations derived by combining the evidence with enterprise architecture, risk, service-management and transformation practice.
- **Review assurance:** independent AI-assisted reviews are treated as challenge inputs, not authoritative sources; material claims are verified before incorporation, and human SME review remains appropriate before regulated reliance.