PROTECT

Stopping harm at the boundary of every AI system

The 15 Guardrails of AI — Pillar 2 of 5

A Vertex11 Whitepaper

7 guardrails · 41 controls · 28 key

G2 Data Classification & Leakage Control (7 controls, 5 key)

G3 Identity & Access Controls (5 controls, 4 key)

G4 Agent Intent & Mandate Control (5 controls, 4 key)

G5 AI System & Supply-Chain Security (10 controls, 6 key)

G6 Prompt & Input Validation (5 controls, 3 key)

G7 Output Validation & Quality Control (5 controls, 3 key)

G9 Tool & Agent Permission Control (4 controls, 3 key)

Validated against NIST AI RMF 1.0, the EU AI Act, ETSI TS 104 223, and ISO/IEC 42001:2023

Executive summary

Protect is where AI governance does its heaviest lifting. Every guardrail in this pillar sits at a boundary of the AI system — around the data that enters and leaves it, the identities that operate it, the instructions it receives, the tools it can reach, and the supply chain it depends on. Get the perimeter right and the majority of real-world AI incidents never happen. Get it wrong and no amount of oversight, assurance or recovery further down the line will fully compensate.

Where Visibility answered what AI you have, Protect answers what stops it going wrong. The pillar contains seven guardrails — from data classification through supply-chain integrity to agent permission control — carrying 41 control statements, 28 of them key, validated against four international standards: NIST AI RMF 1.0, the EU AI Act (as amended by the 2026 Digital Omnibus), ETSI TS 104 223, and ISO/IEC 42001:2023. Every control traces to at least one standard provision.

This paper details the risks each guardrail counters, what the controls cover, and what is required across people, process, and technology to sustain them.

Why protect is the centre of gravity

Detection and recovery are essential, but they operate after something has already gone wrong. Protect grounds everything from the start, reducing how often that happens — and unlike downstream stages, it does so at the lowest unit cost. A control that stops sensitive data from ever reaching a model is far cheaper than the breach response, regulatory penalty, and reputational repair that follow a leak.

The seven guardrails map to the boundaries of an AI system. Data controls (G2) govern *what it knows*. Identity controls (G3) govern *who and what can operate it*. Mandate controls (G4) govern *what an agent is authorised to pursue*. Supply-chain controls (G5) govern *what it is built from*. Input controls (G6) govern *what it is told to do*. Output controls (G7) govern *what it is allowed to produce*. Permission controls (G9) govern *what it can reach when it acts*. A gap in any one boundary is a gap in the perimeter — and attackers, like water, find the lowest point.

The standards validation reinforced this positioning. NIST's MANAGE function assumes controls exist at each interface. The EU AI Act imposes data-governance, transparency and provider obligations that cannot be met without them. ETSI TS 104 223 spells out technical baselines at the model, data and API layers. ISO/IEC 42001 wraps the pillar in policy, roles and dataset-lifecycle expectations. Four standards, one message: the perimeter is where compliance and safety are earned.

The risks that manifest without it

Sensitive-data leakage

Confidential records, source code and personal data are pasted into prompts, absorbed by third-party services, memorised by fine-tuned models, or surfaced through generated output. Under GDPR, the EU AI Act, and sector regulation, each pathway is a distinct liability — and once data has left, it cannot be recalled.

Untraceable access and the confused deputy

Where humans and agents share accounts, or where non-human identities (NHIs) are provisioned without lifecycle, actions cannot be attributed and access cannot be bounded. Agents granted broad credentials become confused deputies — trusted identities tricked into acting on an attacker's behalf. Pilot access is rarely revoked; standing entitlements accumulate.

Function creep in autonomous agents

An agent deployed for one purpose drifts — through emergent behaviour, user coercion, or quiet re-scoping — into adjacent territory it was never risk-assessed for. Without a bound and enforced mandate, the organisation discovers new capabilities only when they cause harm.

Poisoned or opaque supply chain

Models arrive as third-party artefacts with unknown provenance. Training data is scraped from sources that included adversarial content. Weights are modified in the pipeline. Dependencies are patched infrequently or not at all. The AI system is only as trustworthy as the least-known component in its bill of materials.

Prompt injection and jailbreaks

Prompt injection is the defining vulnerability of LLM applications: a hostile instruction hidden in a document, web page, or email hijacks system behaviour without touching its code. Indirect injection through retrieved content is especially dangerous for agents that browse and read. Jailbreaks bypass safety constraints; crafted inputs exfiltrate data the model can access.

Unchecked outputs

Models produce confident, well-written answers that are simply wrong — and a plausible falsehood acted upon is more damaging than an obvious one. Outputs carry bias, generate insecure code, leak training data, or breach compliance requirements. For generative systems the risk extends to unlawful content — including non-consensual imagery and Child Sexual Abuse Material (CSAM) — which the EU AI Act now explicitly requires safe-harbour design against.

Excessive tool and agent permissions

An agent given broad tool access, or a delegation chain that widens authority as it passes between agents, magnifies the impact of any successful compromise. The blast radius of an incident is a direct function of what the agent could reach at the moment it went wrong.

Guardrail 2: Data Classification & Leakage Control

7 controls · 5 key | Pillar: Protect | Applicability: Universal

G2 governs *what the model knows*. Sensitivity labelling is applied before data reaches any model, and enforced through ingestion, retention, output and disposal.

What the controls cover

The controls address the full data life. Classification (G2.1) determines whether and how each dataset may be used in prompts, retrieval, training or fine-tuning. Technical enforcement (G2.2) prevents ingestion above the approved tier and applies DLP and redaction to prompts and outputs. Personal and regulated data is masked, tokenised, or scrubbed before processing wherever raw values are not lawfully required (G2.3). Data sent to third-party models is bound by controls preventing onward use, retention or training (G2.4).

ISO/IEC 42001 widens the aperture. Every dataset — not only training data — is documented for source, selection method, known bias, data rights and prior use, and verified against explicit quality criteria before use (G2.6). Provenance and lineage are tracked across the data's life, and preparation techniques (cleaning, labelling, augmentation, encoding) are recorded per dataset (G2.7). ETSI adds secure disposal on decommissioning or transfer, with the responsible data owner involved (G2.5).

Guardrail 3: Identity & Access Controls

5 controls · 4 key | Pillar: Protect | Applicability: Agentic, Universal

G3 governs *who and what operates the system*. Every human and every AI agent is issued a verifiable, attestable, non-shared identity, and access is bounded to what the task actually needs.

What the controls cover

Identity is not shared and not implicit. Every human and agent gets a verifiable identity at creation (G3.1). Access to models, data, tools and agent capabilities is governed by role — or attribute-based policy on least-privilege terms (G3.2). Non-human identities are lifecycle-managed — provisioned, rotated, de-provisioned — with the same rigour as human accounts (G3.3), closing the standing-credentials gap that turns agents into confused deputies. Authentication strength is proportionate to resource sensitivity and action consequence (G3.4).

ETSI adds environmental separation: development and model-tuning occur in dedicated environments, technically separated from production and bound to least-privilege access (G3.5). This is the control that prevents a tuning experiment from becoming a production incident.

Guardrail 4: Agent Intent & Mandate Control

5 controls · 4 key | Pillar: Protect | Applicability: Agentic, Universal

G4 governs *what the agent is authorised to pursue*. It exists because agents, unlike traditional software, can generalise — and generalisation without a boundary becomes function creep.

What the controls cover

Each agent has a declared, approved, documented mandate, risk-assessed against its intended context, before deployment (G4.1). Permitted actions, tools and scope are bound to that mandate and enforced at runtime, not merely stated in policy (G4.2). Attempted out-of-scope action is blocked and logged (G4.3) — the enforcement evidence that supports later audit. Material change to purpose triggers re-approval and re-assessment (G4.4), preventing quiet re-scoping.

ISO/IEC 42001 requires non-functional and responsible-AI requirements — fairness, robustness, privacy, safety targets and per-use-case tolerances — captured and version-controlled alongside functional requirements before build (G4.5). Responsible AI is a design input, not a post-hoc constraint.

Guardrail 5: AI System & Supply-Chain Security

10 controls · 6 key | Pillar: Protect | Applicability: Conditional, Universal

G5 is the deepest guardrail in the pillar. It treats the AI system as an attackable asset — model weights, training data, pipelines, third-party components and dependencies — and applies security engineering to each layer.

What the controls cover

Integrity and provenance. Weights, pipelines and artefacts are integrity-protected against unauthorised modification, with cryptographic hashes or signatures shared with downstream consumers. Provenance is established and recorded for all third-party and open-source models and components before use; for publicly-sourced training data, records include source and acquisition date to support poisoning forensics.

Adversarial robustness and threat modelling. Systems are assessed for adversarial robustness — evasion, poisoning, extraction, model-layer injection — proportionate to risk. AI-specific threat modelling, covering data poisoning, model inversion, membership inference, extraction and superfluous-capability surface, is performed and reviewed on material change.

Data governance for high-risk systems. Under EU AI Act Article 10, training, validation and test datasets are governed for relevance, representativeness, accuracy and freedom from harmful bias; special-category data used for bias detection or correction meets a strict-necessity test.

Supply chain, patching and disclosure. Supply-chain risk across model, data, dependency and provider is assessed and monitored across the lifecycle. Security patches for AI components are tracked and delivered on a defined cadence, with contingency plans for components that can no longer be patched. A vulnerability disclosure policy for AI systems is published and kept current.

Recoverability and disposal. Where poisoning, tampering or corruption is detected, affected models and datasets can be rebuilt to a verified known-good state. Models, weights and configuration are securely disposed of on decommissioning or transfer.

Guardrail 6: Prompt & Input Validation

5 controls · 3 key | Pillar: Protect | Applicability: Agentic, Universal

G6 addresses the primary attack surface. Every input the model consumes — direct user prompt, retrieved document, tool output, upstream agent message — is treated as untrusted until validated.

What the controls cover

Inputs, direct and indirect, are validated and sanitised against injection. Untrusted content is segregated from instruction context so it cannot be interpreted as commands — the hard boundary between data and instructions that neutralises indirect injection through retrieved content. Input rigour scales with the

agent's authority: an agent that can spend money or send email needs a stricter input path than one that summarises text. Injection attempts are detected, blocked and logged, feeding both real-time defence and the assurance evidence base at G11.

ETSI adds abuse-detection at externally exposed APIs: rate-limiting and abuse controls proportionate to extraction, reverse-engineering and poisoning risk.

Guardrail 7: Output Validation & Quality Control

5 controls · 3 key | Pillar: Protect | Applicability: Generative, Universal

G7 governs *what the model is allowed to produce*. Outputs are validated for accuracy, safety, policy compliance and format before they reach a user or trigger a downstream action.

What the controls cover

Outputs are validated for quality, safety and policy compliance before release. They are checked for leakage of sensitive or personal data prior to release. Generated content carries provenance or watermarking where required for disclosure or downstream trust. High-consequence outputs are gated — by human review or automated guardrail — before they trigger action; this is the control that prevents a plausible falsehood from becoming an executed decision.

The EU AI Act (as amended by the 2026 Digital Omnibus) obliges generative systems to screen for unlawful and seriously harmful content, including a safe-harbour design against non-consensual intimate imagery and CSAM where the system can synthesise images, audio or video.

Guardrail 9: Tool & Agent Permission Control

4 controls · 3 key | Pillar: Protect | Applicability: Agentic

G9 bounds the blast radius. It answers the question that determines the severity of any agent incident: *when this agent goes wrong, what could it reach?*

What the controls cover

Agents and tools receive the minimum permissions, tools, data and execution rights needed for the approved task. Permissions are enforced continuously at runtime and re-evaluated, not granted once and assumed. Delegation chains preserve least-privilege — authority narrows, never widens, as it passes between agents, closing the multi-hop escalation path that turns a low-privilege prompt injection into a high-privilege action. Tool and agent permissions are inventoried, reviewed and revocable.

Building the capability: people, process, technology

People

Role	Responsibility
AI governance lead	Owns the Protect pillar end to end; accountable for perimeter coverage across guardrails, and for the standards-mapped evidence base.
Data protection / privacy office	Sets classification policy, adjudicates edge cases, and owns the lawful-basis analysis for AI processing.
Identity & access management	Owns human and non-human identity lifecycle, including agent identities and their credentials.
Application & platform engineering	Implement input and output validation, model integrity, and supply-chain security at build and deploy time.
Security & red team	Perform AI-specific threat modelling, adversarial testing, injection probing, and vulnerability disclosure handling.
Business-unit / system owners	Declare and maintain the mandate for each agent; approve permissions to least privilege; certify access reviews.
Procurement & vendor management	Assess supply-chain risk, contract for onward-use restrictions, and manage third-party model provenance.

Accountability is the decisive factor. A perimeter with unclear ownership degrades at every boundary in parallel — data owners assume identity owners are handling it, both assume engineering owns validation, and every boundary quietly weakens. Within the Protect pillar, this is accountability at the execution level, which highlights the distribution of accountability for AI systems also required at the macro level.

Process

Classification at intake. Every dataset entering scope is classified, and classification determines eligibility for prompts, retrieval, training and fine-tuning. Data minimisation is the default.

Agent mandate approval. Every agent has a written, risk-assessed mandate before deployment. Material change triggers re-approval. Non-functional and responsible-AI requirements are captured alongside functional ones.

Least-privilege by design. Access, tool and delegation scopes are set at the minimum needed for the approved task, reviewed on cadence and on change, and revoked promptly on role change or system retirement.

Injection testing in the SDLC. Direct and indirect injection tests are part of pre-deployment validation, not a post-incident lesson. Retrieval sources are treated as untrusted content boundaries.

Output gating proportionate to consequence. High-consequence outputs are gated by human review or automated guardrail before they trigger action; grounding and citation are required where accuracy is material.

Supply-chain due diligence. Third-party models and datasets carry recorded provenance; dependencies are patched on cadence; a vulnerability disclosure policy is published; contingency exists for components that cannot be patched.

Secure disposal. Data, models and configuration are securely disposed of on decommissioning or transfer, with the responsible owner involved.

Technology

DLP integrated with AI entry points — prompt-side and output-side, with classification-aware policy rather than pattern-matching alone.

Identity infrastructure that treats agents as first-class principals — non-human identity management with rotation, attestation and revocation, and a hard boundary between development and production environments.

Runtime mandate enforcement — policy engines that evaluate each action against the agent's declared mandate at execution time, blocking and logging out-of-scope attempts.

Model integrity and provenance tooling — signing, hashing, and SBOM-equivalent records for models and their training data.

Input isolation and sanitisation — retrieval and tool-output content segregated from instruction context, with system-level constraints that prevent observed content from overriding operating rules.

Output validation stack — safety and toxicity filters, grounding against trusted sources, schema validation, and secondary checks for critical outputs.

Permission and delegation control — runtime policy that narrows scope through delegation chains and revokes on completion, feeding the tool and agent permission inventory.

How the standards shaped this pillar

NIST AI RMF 1.0 anchors the risk-informed proportionality that runs through the pillar — controls scale with the mapped risk of each use case, not applied uniformly.

The EU AI Act (as amended by the 2026 Digital Omnibus) adds obligations that are conditions of market access: Article 10 data governance for high-risk systems, and the generative safe-harbour design in Article 5(new)/50. These are not aspirational; without them, in-scope systems cannot be placed on the market.

ETSI TS 104 223 closes the operational gaps that most enterprises carry into production — environmental separation for tuning, threat modelling coverage, patching cadence, vulnerability disclosure, abuse-detection at exposed APIs, secure rebuild after poisoning, and secure disposal.

ISO/IEC 42001 wraps the pillar in dataset-lifecycle rigour (source, quality, provenance and preparation documented for every dataset, not only training data) and requires responsible-AI requirements to be captured as first-class design inputs.

The net effect: the Protect pillar covers data governance, identity and mandate control, supply-chain integrity, input and output validation, and permission bounding to the expectations of all four standards, from a single set of controls.

A maturity path

Level 1 — Ad hoc perimeter. Classification is informal, DLP is not integrated with AI entry points, agents share credentials, mandate is implicit, supply chain is untracked. Most organisations sit here for at least one guardrail.

Level 2 — Documented boundaries. A classification scheme exists and is enforced at ingestion for the most sensitive tiers. Non-human identities are provisioned centrally. Agents have written mandates. Input and output validation exists for the highest-risk systems. Supply-chain provenance is recorded for named third-party models.

Level 3 — Enforced and tested. Runtime enforcement of mandate and permissions. Injection testing in the SDLC. Third-party model provenance recorded across the estate. Recoverability from poisoning tested. A vulnerability disclosure policy is published.

Level 4 — Continuously assured. Classification-aware DLP across all AI entry points. Non-human identity lifecycle fully automated. Runtime policy enforcement across mandate, tool and delegation scope. AI-specific threat modelling on cadence. Supply-chain integrity verified continuously. Output validation proportionate to consequence, embedded in the deployment pipeline.

The goal is not uniform maturity across all seven guardrails on day one, but a defensible baseline at each boundary — with depth added where the combination of autonomy, data sensitivity and decision impact is greatest.

Connecting to the rest of the framework

Protect consumes the risk rating and inventory from Visibility (G1) — the perimeter is applied proportionately to what each system is, does, and touches. Its logging obligations feed the tamper-evident record and monitoring stream at G10.

Downstream, Protect determines the residual risk that Oversight (G8, G15) must manage in-flight. It reduces the surface that Assurance (G11, G12) must test. It shrinks the incident population that Resilience & Recovery (G13, G14, G15) must recover from. A well-defended perimeter is not a substitute for the later stages, but it is what makes them tractable.

About this series

This is Pillar 2 of The 15 Guardrails of AI, published by Vertex11. The series covers the full framework across five whitepapers — Visibility, Protect, Oversight, Assurance, and Resilience & Recovery — each aligned to the guardrails, controls, and standards mapping in the V11 AI Governance Framework.

To pressure-test the perimeter around your AI systems, get in touch.