VIANORDIS POSITION PAPER 01

The AI Control Gap

Why approving a model is not approving an action

From Shadow Execution to Governed Intelligence

Version	Draft v0.5.4
Status	Open Review Edition
Date	26 August 2026
Author	Yves Frédéric N'Soussoula, Founder & CEO
Publisher	GoSec Cloud OÜ, Tallinn, Estonia
Licence	Creative Commons Attribution 4.0 International
Web	www.vianordis.eu

Open Review. This document is published for criticism. Corrections of fact are applied in the next edition and listed; substantive critique is credited unless the contributor prefers otherwise. Contact: www.vianordis.eu/en/contact · Law as at 26 August 2026. Not legal advice.

Contents

On one page 4

How to read this set 6

Document status, audience and scope 7

- Audience 7
- Core thesis 7
- Evidence status — read this before Chapter 1 7
- Normative language 7

Executive summary 8

1. When an answer becomes an action 10

- 1.1 Governance mostly stopped at model access 10
- 1.2 A prompt is not a mandate 10
- 1.3 What existing controls actually cover 10
- 1.4 An approval is not yet a transaction commitment 11

2. From Shadow AI to Shadow Execution 12

- 2.1 Shadow AI describes only part of the risk 12
- 2.2 Definition 12
- 2.3 Observable forms 12
- 2.4 Impact tiers: not every AI use needs the same depth of control 13
- 2.5 Two dimensions of risk: impact and provenance assurance 13

3. What this architecture does not solve 15

- 3.1 The general list 15
- 3.2 Input manipulation: the limit that matters most 15
- 3.3 The control plane is a single point of failure 16
- 3.4 Cost, latency and brownfield reality 16
- 3.5 Employee representation, co-determination and the dual-use problem 17
- 3.6 State changes after approval: the TOCTOU limit 17
- 3.7 Signatures, time and long-term evidence have a lifecycle 18
- 3.8 What the control plane does answer 18

4. Governed Intelligence: a working definition 19

- 4.1 From governance documents to executable governance 19
- 4.2 Definition 19
- 4.3 The eight elements 19
- 4.4 Twelve control invariants 20

5. Regulatory convergence becomes an architecture problem 22

- 5.1 No single regulation requires a control plane 22
- 5.2 EU AI Act: what applies when 22
- 5.3 Germany: NIS2 and the BSI Act — management liability is already live 24
- 5.4 DORA: responsibility survives outsourcing 25
- 5.5 GDPR: purpose and minimisation have to hold at execution time 26

5.6 Two instruments often left out 26
5.7 Shared technical objectives 27
6. Conclusion **28**
Annex A — Control capabilities and regulatory objectives **29**
Annex B — Claims this architecture must not make **31**
Annex C — Primary sources **32**
Annex D — Version history **33**
Imprint and contact **37**

> **Publisher's disclosure.** This paper is published by GoSec Cloud OÜ (Tallinn, Estonia), which develops and operates **Vianordis**, an implementation of the architecture class described here. We have a commercial interest in the argument. We have tried to write a paper that survives being read by someone who does not. Chapter 3 states what this architecture does *not* solve; the companion specification states what already exists and does part of the job; the implementation status page states exactly what is built and what is not. If you find those insufficient, tell us — the contact route is at the end.
>
> **Legal status.** Law as at **26 August 2026**. This is a strategic and technical position paper, not legal advice. Deadlines, interpretations and secondary sources change; the linked official texts govern. Nothing here replaces legal classification of a specific AI system, a data protection impact assessment, a conformity assessment, an information security review, or sector-specific analysis.

On one page

What changed. AI no longer only produces text. It reads mailboxes, joins records across systems, prepares transactions, operates applications and triggers actions. The output of a language model has become a business event.

Where the gap is. Many organisations now control *which model* may be used. This paper's central hypothesis is that materially fewer can demonstrate, for a specific action, *who authorised it, on whose behalf, under which mandate, under which policy version, from which sources its decisive parameters were derived, and whether the thing approved under a given system state is the thing that executed.* Access is not authority, and approval is not a guarantee that the underlying facts were true or still current.

Why this year. Four obligations bear on this, three of them already in force:

- **§ 38(2) BSIG (Germany, in force since 6 December 2025)** — management bodies of in-scope essential and important entities have implementation and supervision duties; a culpable breach can create personal liability toward their own organisation under the company-law rules applicable to the entity.
- **DORA (since 17 January 2025)** — the management body defines, approves, oversees and is accountable for ICT risk arrangements, and the financial entity remains responsible when ICT services are outsourced.
- **EU AI Act** — broadly applicable since 2 August 2026; the core high-risk obligations now apply from **2 December 2027** for Article 6(2) / Annex III systems and **2 August 2028** for Article 6(1) / Annex I systems, following the Digital Omnibus on AI. That is implementation time, not a reason to postpone architecture work.
- **Cyber Resilience Act — Article 14 from 11 September 2026**, the one that starts next month. Manufacturers of in-scope products with digital elements must report actively exploited vulnerabilities on a 24 h / 72 h / final-report-within-14-days-after-a-fix-is-available sequence, and severe incidents on a 24 h / 72 h / one-month sequence. This affects many suppliers of distributable, embedded and on-premises software, but not every pure SaaS provider.

What to do. Inventory AI *actions*, not only AI models. Pilot one bounded tier 3 or tier 4 process. Give it a distinct agent identity, a scoped mandate, a versioned policy decision, a governed action envelope, a bound approval, an execution receipt and a correlated evidence record. Then measure reconstruction time, bypass coverage, operational cost, and — separately, never blended — preventive stale-state rejection where the target system can enforce it and divergence-detection coverage and latency where it cannot, before extending the pattern.

Three questions for IT leadership.

1. For an action an AI agent took last week, can we reconstruct the human or organisational principal, purpose, policy version and source of every impact-determining parameter — from records, without asking the team that built it?
2. If a person approved a payment, is the approval bound not only to amount, payee and invoice, but also to the material system state under which it was approved — and when that state changes, is the payment **prevented**, or only **detected**

afterwards, and within how long?

3. Have all alternative paths to a tier 4 or tier 5 target operation outside the enforcement point been inventoried, and will reconciliation detect divergence from the governed evidence path within a declared interval — and has the safe-failure behaviour been tested?

What this paper does not claim. That regulation mandates any particular product. That a control plane makes an organisation compliant. That signatures prove a decision was correct. That an approval remains valid after the world changes. That any architecture defends against a fully compromised administrative domain or against an agent acting with valid authority on attacker-supplied facts. See Chapter 3.

How to read this set

Version 0.4 of this paper ran to roughly 30,000 words and contained three documents in one envelope: an argument, a technical profile and an adoption programme. Each audience had to read past the other two. Version 0.5 splits them.

Document	What it contains	Who it is for
Position Paper 01 — The AI Control Gap (this document, v0.5.4)	The problem, its limits, the working definition, and the regulatory case with dates	Boards, CIOs/CISOs, compliance, data protection, works councils
GAIP Candidate Specification 0.4	The normative technical profile: threat model, roles, envelope data model, digests, signer matrix, processing rules, evidence requirements and negative tests	Architects, implementers, assessors
Adoption & Measurement Guide 0.1.4	A worked example, the six-phase adoption path, brownfield enforcement modes, the maturity model, and the measurement protocol	Programme owners, pilot teams
Implementation status (web page, rolling)	What Vianordis actually is, application by application, and what has not been independently verified	Prospective adopters, the community

The three documents version independently. A change to the technical profile no longer requires a new edition of the position paper, and the reverse. Where this paper needs a technical detail it names the companion document rather than reproducing it.

What moved out of this edition, and where it went. The architecture of a runtime control plane, the relationship to prior art, the trust-boundary analysis and the candidate interchange profile are now in the *GAIP Candidate Specification*. The worked example, the adoption path, the maturity model and the measurement protocol are in the *Adoption & Measurement Guide*. The Vianordis implementation status is on the release-status page, because a printed status ages inside the document. Chapter 11 of v0.4, on open source and institutional stewardship, is held for separate publication; its subject is the licence and foundation model, not the control gap.

Document status, audience and scope

Audience

This paper is written for people who will have to answer for an AI-initiated action after it happened: executive boards and managing directors; CIOs, CTOs, CISOs and CDOs; IT and enterprise architects; data protection, compliance and risk owners; owners of AI platforms and automation; works councils and employee representatives (see §3.5, which is written for them specifically); and developers, maintainers and open-source contributors.

Core thesis

> **Enterprises do not fail to scale AI only because of model quality or regulation. They fail because organisational authority and technical execution are not connected.**

This thesis is falsifiable. The *Adoption & Measurement Guide* states four measurements that would confirm, weaken or falsify material parts of it. We do not yet have those measurements at scale. Saying so is part of the argument.

Evidence status — read this before Chapter 1

This is a **position paper**, and its central claim is currently supported by architecture and reasoning, not by published field data. We are explicit about that for two reasons. First, an unmarked assertion dressed as a finding is exactly the kind of governance artefact this paper argues against. Second, the measurement protocol in the companion guide is the mechanism we are asking the industry to help gather.

Where this paper cites law, it cites the official text. Market-wide statements are framed as hypotheses unless supported by published evidence. Readers who find the lack of field data unsatisfying are reading it correctly.

Normative language

> **Release status.** Open Review position paper. The central thesis is an architectural hypothesis supported by reasoning, legal analysis and prior art — **not yet by cross-organisational field measurements**. The companion GAIP document is an incomplete candidate specification and defines no certification or conformance scheme.

This document is explanatory and **non-normative** throughout. It contains no MUST, MUST NOT or SHOULD requirements. Those belong to the *GAIP Candidate Specification*, where they describe proposed conformance requirements for a candidate profile — not statements of law, and not a certification scheme. That specification withdrew the conformance levels present in v0.4 of this paper; the reasons are stated there.

Executive summary

The enterprise conversation about AI is changing shape. The first phase was chatbots, text generation and personal assistants. The next is systems and agents that reach into data, prepare business transactions, operate applications and trigger actions.

The decisive question moves with it:

> **Not only: which AI may we use? But: who authorises the AI when it acts on our behalf?**

Many organisations now have AI policies, approved model catalogues, data protection rules, API gateways and technical logging. All of these are necessary. Individually, none of them closes the chain between a business intention and the action that actually executed.

A valid API key proves that a system was technically able to reach an application. It does not prove that the specific action was organisationally mandated, purpose-bound, proportionate, and — where required — approved by a competent human.

> **Access is not authority.**

From that separation comes the **AI control gap**. Organisations may be able to identify the model or technical account that triggered a call. The hypothesis tested by this paper is that, for a material share of actions, they cannot demonstrate from existing records: which natural or legal person stood behind it; on whose behalf the agent acted; what mandate it actually held; which data and systems were permissible for that purpose; which policy version applied at the moment of decision; whether human approval was required; whether the approved action is the action that executed; and what verifiable event chain remains.

This paper calls AI-assisted execution that cannot be reconstructed **Shadow Execution**. It can arise inside a fully approved environment, with an approved model, run by an approved team. The problem is not the model. It is the missing link between identity, mandate, permission, policy, approval, execution and evidence.

Shadow Execution is not a new phenomenon. It is the confused-deputy and delegation problem, known since the 1980s, and it applies equally to RPA bots, cron jobs and CI/CD pipelines. What is new is the scale: agents generate delegation chains at a frequency and branching depth that makes manual reconstruction impractical.

Regulation raises the cost of not solving it — though not in the way most summaries suggest. The AI Act's high-risk obligations were **deferred** by the Digital Omnibus on AI to December 2027 and August 2028. The obligations that bite *today* are elsewhere: personal management liability under German § 38(2) BSIG, management-body accountability under DORA, and the Cyber Resilience Act's 24-hour vulnerability reporting clock from 11 September 2026. Chapter 5 sets out all four, with dates.

Governed Intelligence is the operating state in which an AI-assisted action does not execute merely because it is technically reachable, but passes an enforceable, reviewable and revocable control chain:

> **Intent → Identity → Mandate → Context → Policy → Approval → Execution → Evidence**

Governance then stops being reconstructed only afterwards in reports. It becomes a precondition of the action.

The companion specification makes that proposition concrete. The chain is carried in a **Governed Action Envelope** that binds the proposed action, parameter provenance, mandate, policy decision, approval, expiry and material execution preconditions. At commit time the enforcement point revalidates revocation and state. If the target state has changed, the approval is stale. What happens next depends on the target system, and the distinction is not cosmetic: where the target can enforce a conditional commit or hold a lock, the transaction is prevented from committing and must be re-evaluated or re-

approved; where it cannot — and whether it can is a property of the specific API and resource, to be tested rather than assumed — the action may still commit, and the divergence is detected and escalated afterwards within a declared interval. The resulting execution receipt is joined with the decision record in the evidence layer.

This does not make distributed transactions perfectly atomic. It makes the residual gap explicit: an action can be approved correctly and still become invalid before execution. That time-of-check/time-of-use problem is addressed in §3.6.

Almost nothing in this architecture is new, and the companion specification says so at length. The decision/enforcement split and the PDP/PEP vocabulary come from the IETF policy framework work of 2000–2001, with antecedents in ISO/IEC 10181-3; XACML adopted that vocabulary and added PIP and PAP. Delegation semantics exist in OAuth 2.0 Token Exchange (RFC 8693), macaroons and GNAP. Workload identity exists in SPIFFE/SPIRE. Tamper-evident logs come from Certificate Transparency. Supply-chain integrity comes from SLSA, in-toto and Sigstore. Two mechanisms are worth naming precisely because they are closest to the claim: **OAuth 2.0 Rich Authorization Requests (RFC 9396)** already carries transaction-specific authorisation detail such as amount, currency and creditor; and **PSD2 dynamic linking** (Commission Delegated Regulation (EU) 2018/389, Art. 5(1)) has required since 2019 that a payment authentication code be specific to the amount and the payee, and be invalidated when either changes. Binding an authorisation to a concrete transaction is therefore not new — it is settled European payments law. Recent agent-governance work goes further still. Pretending otherwise would be the fastest way to lose a technical reader.

The contribution claimed is accordingly narrow: a candidate *composition* profile that combines attested parameter provenance, binding to a signed rendering artefact — and, where the approver's own client renders and signs it, to what the approver actually saw — mutable target-state preconditions, full actor-chain attribution, cumulative delegation limits, write-ahead evidence and cross-system decision-and-execution correlation. It claims no priority over execution-time authorisation, transaction dynamic linking or governance-receipt architectures.

Chapter 3 states what none of this fixes — including the case that matters most: if a model derives action parameters from an attacker-controlled document, a correctly bound approval binds to falsified values. The chain is formally intact and materially worthless. Any architecture that does not say this out loud should not be trusted with the rest.

Vianordis is our implementation of this architecture class: a model-independent control layer for people, agents, applications and business processes. It is neither a language model nor a wrapper around one. As of this edition it is in **Controlled Beta**, by invitation, with no public source and no completed certification. The release-status page lists every application, its status, and what has not been done yet.

---

1. When an answer becomes an action

1.1 Governance mostly stopped at model access

The first wave of enterprise AI governance was about tool selection: may this AI service be used at all; what data may be entered into a model; which vendor satisfies our data protection requirements; which models are centrally approved; may staff use personal accounts; which outputs must be labelled.

These questions remain relevant. They are questions about **access to AI**.

Agents add a second layer. An agent can read mail and documents, join CRM, ERP and HR records, create tasks and appointments, prepare purchase orders, create accounts and permissions, change configurations, publish content, prepare payments and contract steps, and drive external systems through APIs.

At that point a linguistic output has become a business event.

> **A model answers a question. An agent changes a state.**

Architecturally, this is the whole difference.

1.2 A prompt is not a mandate

Consider: *"Check this invoice and pay it if everything is correct."*

The sentence expresses intent. It does not answer:

- Who is entitled to issue this instruction?
- Does that person represent only their own cost centre, or the whole company?
- May the agent only prepare a payment, or also release it?
- What amount threshold applies?
- Which suppliers may be processed?
- Is dual approval required?
- Which data may be transmitted to the selected model?
- What happens if the agent interprets the instruction differently?
- How is it ensured that the action executed after approval is the action approved?

Natural language expresses intent. It is not by itself sufficient technical authorisation.

1.3 What existing controls actually cover

The table below is deliberately written in the strongest reasonable form of each existing control. Earlier drafts of this paper understated them, which made the argument easier to write and easier to dismiss.

Existing control	What the mechanism actually covers today	What is typically still outside its model
Model catalogue	Which models and vendors are approved, under which data classes	What a specific action may do in a specific business transaction
IAM and SSO, including delegation flows such as OAuth 2.0 Token Exchange (RFC 8693), PIM/JIT elevation, workload identity federation and Continuous Access Evaluation	Authentication, delegated technical authorisation, actor claims, session revocation	Purpose binding, monetary and risk limits, and revocation of a single *assignment* rather than a session

Existing control	What the mechanism actually covers today	What is typically still outside its model
Roles and groups	Which application and which coarse operations are reachable	Which purpose the entitlement was granted for, and which single action it justifies
API gateway	Scopes, claims, actor claims, rate limits, external PDP calls	Whether the resulting business act was organisationally mandated and proportionate
Data loss prevention	Which data classes may leave which boundary	Whether use of that data is lawful *for this purpose*
Prompt and output filters	Which inputs and outputs are unwanted	Which real-world action may result from an accepted output
Workflow approval (SAP, Coupa, Workday and comparable)	Approval bound to a document object inside that system	Approval bound to an action executed by an agent *across* systems, including parameter provenance
SIEM and logging	Which events were recorded, per system	Which policy version and which mandate justified the act, correlated across systems
GRC documentation, ISO/IEC 42001, ISO/IEC 27001	Which rules, controls and management processes were decided and audited	Whether those rules were technically enforced at the moment of execution

None of these controls is redundant. Several of them do more than this paper's earlier drafts credited. The gap is not that they are weak; it is that they were designed for different questions, sit in different systems, and do not compose into one authorisation and evidence chain.

1.4 An approval is not yet a transaction commitment

The action described to an approver is evaluated at a point in time. The target system executes later. Between those moments, the relevant state may change: a supplier may be blocked; a bank account may be replaced; a purchase order may be closed or amended; an entitlement may be revoked; a sanctions-screening result may change; a policy or risk threshold may be updated; another process may consume the budget the action assumed was available.

This is a classic **time-of-check/time-of-use (TOCTOU)** problem. Binding an approval only to the visible parameters is therefore insufficient. A governed action needs two distinct commitments:

1. an **action commitment** — a digest of the operation and the parameters the approver saw; and
2. a **precondition commitment** — a digest or version reference for the material state that made the operation permissible.

At execution time the action broker must verify both. Where the target system supports conditional writes, optimistic concurrency, version identifiers or compare-and-set semantics, the broker should use them. Where it does not, what is available is weaker and must be named as such. For consequential actions the companion specification requires an authoritative broker-side intent ledger with a declared reconciliation interval: the action commits, and a divergence is detected and escalated afterwards rather than prevented. A read-then-write immediately before execution is permitted only for lower-impact actions, because the race it leaves is not acceptable where the action is consequential. The specification defines these assurance modes and is candid that the strongest of them is unavailable against several major enterprise targets.

The distinction matters because a perfectly signed approval can authorise an action that is no longer permissible. **Cryptographic binding proves equivalence to what was approved; it does not prove that the approval is still fresh.**

2. From Shadow AI to Shadow Execution

2.1 Shadow AI describes only part of the risk

Shadow AI usually means the use of unapproved AI services: personal accounts, unknown browser extensions, unvetted models. It is partly addressable through model catalogues, network rules, contracts, training and central access.

The next class of risk arises **inside** the approved environment. An officially sanctioned agent can operate with an over-privileged service account; inherit the requesting user's entitlements and keep them; chain tools whose combined effect was never assessed; carry an instruction beyond its original purpose; pass data to a different model or provider; apply an approval to an action other than the one presented; keep acting after a role change or a leaver event; and produce logs across systems that cannot be correlated afterwards.

Shadow AI and Shadow Execution are **not mutually exclusive**, and one is not simply the successor of the other. Shadow AI is a possible cause of Shadow Execution, not a necessary one. Unapproved tools produce unreconstructable actions; so do approved ones.

2.2 Definition

The definition below is deliberately written so that it can be evaluated by an auditor **without reference to any particular product or vendor**:

> **Shadow Execution** is an AI-assisted action for which an auditor cannot, within a defined time budget and from the organisation's existing records, reconstruct (a) the human or organisational principal on whose behalf it was taken, (b) the purpose and scope it was permitted under, (c) which rule set the organisation can show was actually applied at the moment of decision, and (d) whether any approval on record can be tied, by any means, to the values that actually executed.

Shadow Execution does not imply that an action was unlawful or harmful. It means the organisation cannot establish its legitimacy or its course of events.

The risk is therefore not only a wrong model answer. It is that a *correct or plausible* answer becomes an insufficiently authorised action.

On novelty. We make no claim to have discovered a new class of risk. This is the delegation and confused-deputy problem described by Norm Hardy in 1988, and it applies to RPA bots, scheduled jobs and CI/CD pipelines exactly as it applies to agents. Substitute "script-driven action" for "AI-assisted action" in the definition above and it remains true. What agents change is scale and opacity: delegation chains multiply, branch, and are generated at machine speed by a component whose reasoning is not directly inspectable. Manual reconstruction stops being practical somewhere on that curve. That is a quantitative change with qualitative consequences, and it is the honest version of the claim.

2.3 Observable forms

Each of the following is stated as an **observable configuration**, not a documented incident. We are not aware of published, attributable case studies for all of them, and we say so rather than implying otherwise.

Over-privileged technical identities. An agent uses a service account whose permissions were sized for integration convenience rather than for the individual task.

Unbounded delegation. A user signs in once; the agent receives a long-lived token and continues to act later, in a changed context.

Unbound approvals. A human confirms "the payment" without the approval being technically bound to amount, payee, invoice reference and the specific transaction.

Tool chains without combined assessment. Every individual tool is approved. Mail access plus document analysis plus ERP write plus external communication produces an effect that was never assessed as a whole.

Retrospective reconstruction. The organisation holds several logs but cannot reliably join which user instruction produced which model decision, which policy evaluation and which system action.

Policy drift. An agent was approved under an earlier policy. The rule changed; running mandates and agent instances were never re-evaluated.

Trust inheritance through integrations. An approved business application calls an agent that calls further services. The original approval is passed implicitly across system boundaries, with no delegation step visible or bounded.

2.4 Impact tiers: not every AI use needs the same depth of control

This taxonomy is introduced here rather than late in the paper, because Chapter 3 and both companion documents depend on it.

Tier	Effect class of the action	Example	Typical control
0	Inform	Summarise a document	Data and model policy
1	Recommend	Produce a proposed course of action	Labelling, sources, review
2	Prepare	Create a draft email, contract or booking	Mandate, draft marking
3	Execute reversibly	Create a calendar entry, update an internal ticket	Policy, bounded mandate, logging
4	Execute consequentially	Payment above a defined threshold, publication, permission change	Bound human approval
5	Critical or irreversible	Production shutdown, clinical or personnel decision	Strict oversight, multi-party approval, or prohibition

The tier is an attribute of **the action**, not of the model. It is also value-sensitive: a payment below a defined threshold, on a full three-way match, may reasonably sit at tier 3, while the same operation above that threshold sits at tier 4. Where an organisation draws that line is a policy decision. The same model can operate at tier 0 for a document summary and at tier 5 when changing production parameters.

This is an internal architecture and control instrument. It is **not** the legal risk classification of the EU AI Act, and it should not be presented to an auditor as such.

A control point that follows from this and is easy to get wrong: in the companion specification, the tier is **not** declared by the component requesting the action. A requester that could set its own tier could exempt itself from approval, segregation of duties and the no-bypass rule while remaining formally conformant. The tier is derived by the policy decision point from a signed action-type registry and recomputed at the enforcement point.

2.5 Two dimensions of risk: impact and provenance assurance

The impact tier describes what an action can do. It does not describe how trustworthy the facts behind the action are. A second axis is needed.

Provenance assurance classes apply to impact-determining parameters:

Class	Meaning	Typical example
P0 — independently verified	The value was checked through a second, organisationally independent path against an authoritative source	Bank account confirmed through supplier onboarding and call-back verification
P1 — authoritative system state	The value was read from a designated system of record with an authenticated version or state reference	Approved supplier master record, current employee directory record

Class	Meaning	Typical example
P2 — authenticated assertion	The value was entered or asserted by an authenticated person or service, but not independently verified	Cost centre supplied by the requester
P3 — derived from untrusted content	A model or parser derived the value from mail, documents, web content, tool descriptions or other attacker-influenceable input	IBAN extracted from an inbound invoice
P4 — mixed, unknown or untraceable	The value has multiple unresolved sources, an unknown derivation path, or no reconstructable provenance	Parameter assembled by a chained agent with incomplete telemetry

These classes are not legal categories. They are an engineering control for deciding when automation must stop.

A conservative default matrix is:

Action tier	P0–P1	P2	P3	P4
0–1	Permit under data/model policy	Permit with attribution	Permit with attribution	Permit only if the information purpose tolerates uncertainty
2	Permit	Permit with review	Draft only; visually mark provenance	Hold for clarification
3	Permit with normal controls	Permit within bounded thresholds	Verify impact-determining fields before execution	Deny or reduce to tier 2
4	Permit with bound approval and fresh preconditions	Require explicit approval and, where appropriate, second control	Do not execute until critical fields reach P0 or P1	Deny
5	High-assurance process and multi-party control	Exceptional process only	Prohibit automated execution	Prohibit

The matrix is intentionally stricter than many existing workflows. It is a candidate default, not a universal rule. An organisation may choose differently, but the exception should be explicit, versioned and visible in the evidence record.

One point of vocabulary about the matrix above. A parameter for which no provenance assessment was required and none was made does **not** get a class: it carries the status *not assessed*, and the matrix is not consulted for it. That is permitted only at tiers 0 to 2, only where the action type's policy says so, and it is recorded. The taxonomy stays at five classes; "not assessed" is the absence of an assessment, not a sixth degree of assurance.

Two things make this matrix a control rather than a diagram. First, the same principle as for tiers applies: a provenance class asserted by the component that assembled the parameter is worth nothing, so the companion specification requires, for actions at tier 3 and above, attestation by a resolver role that is not the requester, and fails an unattested claim closed to P4. Second, the matrix must be evaluated in the processing path, not merely printed in an annex — an implementation that satisfies this profile stops the action when a prohibited combination appears.

3. What this architecture does not solve

Most papers of this kind place their limitations at the end, where fewer people read them. They are here, before the definition, because a reader who does not know the boundaries cannot evaluate the claim.

3.1 The general list

A control layer does not replace other governance, security and compliance work. In particular, it cannot automatically: determine the legal role of provider or deployer for a specific system; classify an AI system as high-risk or not high-risk in law; replace a data protection impact assessment; guarantee the quality or representativeness of training and input data; prevent discrimination or erroneous model outputs; replace subject-matter plausibility review; train staff; define organisational responsibility; replace an effective information security process; remove supply-chain dependency; replace a statutory conformity assessment or certification; or protect a compromised organisation from the actions of its own highest-privileged administrators.

3.2 Input manipulation: the limit that matters most

This is the most important section of this paper.

A control chain authorises *an action with parameters*. It does not verify that those parameters describe reality.

Consider an agent that reads an incoming invoice, derives supplier, bank details, invoice number and amount, and presents them to a human for approval. The approval is then cryptographically bound to exactly those values, and the action broker executes exactly that transaction.

Now assume the invoice document is attacker-controlled — which, for an inbound invoice mailbox, is the normal case rather than the exotic one. The document contains instructions or formatting designed to steer the extraction: a substituted IBAN, an injected directive, a manipulated line item. The model derives the falsified parameters. The human sees them rendered cleanly in an approval dialogue. The approval binds correctly. The broker executes exactly what was approved.

> **The control chain is formally intact and materially worthless.**

This is indirect prompt injection, and no amount of identity, mandate, policy or evidence architecture prevents it. Related failure modes with the same property: tool-description poisoning, where a malicious tool manifest steers selection; confused-deputy escalation through chained tools; and data exfiltration through a channel the agent is legitimately entitled to use.

What the architecture *can* contribute is narrower, and it should be stated narrowly:

Mitigation	What it actually achieves
Parameter provenance in the evidence record — every decision parameter carries its origin (model-derived from document X / user-entered / read from system of record)	Does not prevent manipulation; makes it reconstructable afterwards and reviewable at approval time
Out-of-band verification of critical fields — bank details checked against supplier master data, not against the document	Prevents the specific highest-value case, at the cost of a second data path
Provenance surfaced in the approval dialogue — the approver sees which values came from the untrusted document	Converts an invisible risk into a visible one; effectiveness depends on the approver actually reading it
Change thresholds on master data — any new or changed payee triggers a second, differently-sourced approval	Standard anti-fraud practice; predates AI and remains the strongest single control
Tier ceilings for untrusted input — actions whose parameters derive wholly from untrusted input are capped at tier 2 (prepare)	Reduces blast radius; reduces automation benefit

The honest summary: a runtime control plane reduces the number of ways an agent can act *without* authority. It does much less about an agent acting *with* authority on false premises. Claims that identity, policy and approval controls alone eliminate input-manipulation risk should therefore be treated with caution.

There is one genuine architectural consequence, and it is why this section exists rather than being a footnote: **evidence without parameter provenance is not evidence of much.** A record proving "person C approved amount X to payee Y" is close to useless in an injection incident. A record proving "amount X and payee Y were extracted by model M version N from document D, hash H, received from sender S, and were not verified against master data" is what an investigation needs. That requirement changes the evidence schema, and it is not in most designs.

3.3 The control plane is a single point of failure

A shared runtime control plane sits in the path of every governed business process. That is the point of it, and it is also its principal operational risk. If it is unavailable, either business stops or it does not — and either answer has a cost.

The design response is graceful degradation, not a claim of perfect availability. If a model becomes unavailable, another may be substituted where policy and data classification allow. If the **authorisation or policy component** fails, the system must not continue acting unconstrained; it should fall to a safe state: read-only access, draft mode, manual handling, or a full execution block for consequential actions. The tier taxonomy in §2.4 determines which actions block and which continue. That mapping must be decided in advance, written down and tested — not discovered during an incident. The companion specification carries the safe-degradation matrix.

An organisation adopting this pattern should treat control-plane availability with the same seriousness it applies to its identity provider, because the failure characteristics are similar.

3.4 Cost, latency and brownfield reality

This paper describes a target architecture. It does not describe what it costs.

- **Latency.** Every governed action adds at least one policy evaluation and, for tiers 4 and 5, a human round trip. Policy evaluation can be single-digit milliseconds; human approval cannot. Processes with tight cycle times need their tier assignments chosen accordingly.
- **Implementation effort.** The adoption path in the companion guide has six phases and no stated durations, because honest durations depend on the estate. A bounded pilot may be achievable within weeks in some estates. Retrofitting hundreds of integrations is a different programme and should not be represented as a scaled version of the same task.
- **Brownfield.** The harder question — how to route two hundred existing integrations through enforcement points without a two-year freeze — is not solved in this edition. The realistic answer involves prioritising by impact tier rather than by system, and accepting a long tail of tier 0–3 integrations that remain outside the enforcement path for some time. We would rather say that than imply a clean sweep. The enforcement modes M0–M4 in the companion guide are the migration ladder, and the guide is explicit that a claim of "no bypass" applies only to the action set actually placed behind an enforcement point.

The absence of field measurements remains a weakness. This edition therefore defines what must be measured without inventing values. For an automatically executed action, the machine-path latency budget is:

> **T_machine = T_identity + T_context + T_policy + T_state-check + T_broker + T_target + T_evidence**

For an approval-bound action:

> **T_end-to-end = T_machine + T_human + T_queue**

The second equation is usually dominated by human and queue time, not cryptography or policy evaluation. The companion guide defines a reporting template for p50, p95 and p99 machine latency, approval latency, false-block rate, exception rate, reconstruction cost, and — reported separately rather than blended — expired-envelope rejection, preventive stale-state rejection where the target can enforce it, and divergence-detection coverage and latency where it cannot. Until measured results are published, no performance claim should be inferred from this paper.

3.5 Employee representation, co-determination and the dual-use problem

This section exists because works councils are named in this paper's audience, and a paper that names them without addressing them is not addressing them.

The evidence architecture described in the companion specification produces a complete, personally attributable, cryptographically protected record of who instructed what, who approved what, and when. That is precisely what makes an action reconstructable. It is also, without any change to the design, an instrument capable of monitoring the conduct and performance of employees.

In Germany, a technical system that is *objectively suitable* for monitoring employee conduct or performance engages the works council's co-determination right under § 87(1) no. 6 BetrVG — regardless of whether the employer intends to use it that way. A personally attributable evidence layer of the kind described here is likely to meet that test, though whether a particular implementation does so depends on its concrete fields, its attribution model, its query functions and its deployment. Comparable rules exist in other jurisdictions with statutory employee representation.

Under the GDPR, the same architecture raises purpose limitation and data minimisation questions in a sharper form than ordinary logging does, because the records are designed to be durable, attributable and cryptographically tamper-evident — which is in tension with erasure and rectification rights. The separation of payload data from integrity evidence is intended to address this; whether it does so adequately is a question for a data protection impact assessment, not for a vendor's whitepaper. A cryptographic digest of personal data can still be personal data where it is linkable or where the controller can resolve it. Hashing is not automatic anonymisation.

What follows for anyone deploying this pattern:

1. **Involve employee representatives at design time, not at rollout.** The architecture decisions that matter — retention per record class, which fields are personally identifiable, who may query the evidence store, whether individual-level reporting is possible at all — are cheap to change early and expensive to change late.
2. **Separate the audit purpose from the performance-management purpose technically, not only in policy.** Access to the evidence store for incident reconstruction and access for any form of individual performance view should be distinct, separately authorised, and separately logged.
3. **Write the works agreement around the tier model.** Tier 0–3 actions generate operational records; tier 4–5 actions generate approval-bound evidence with named individuals. Those are different categories and deserve different treatment.
4. **Expect the question "can this be used to rank us?" and have a technical answer.** "We would not do that" is not one.

A works council reading the companion specification without this section would reasonably conclude that it describes surveillance infrastructure. It describes infrastructure that can be either, and the difference is a set of design decisions that should be made jointly.

3.6 State changes after approval: the TOCTOU limit

An approval can be perfectly bound to the action presented and still be unsafe to execute later. The state that justified the approval may have changed. This is not an AI-specific problem, but agents increase the number of delayed and chained actions for which it appears.

The architecture can reduce the risk by recording material preconditions, assigning the approval a short validity period, revalidating policy and revocation immediately before commit, and using conditional target-system operations where available. It cannot make an external system atomic when that system offers no transaction or concurrency primitive.

This limit is sharper in practice than it looks on paper. Conditional-write, locking and idempotency capabilities vary by API, by resource and by deployment model, and they vary within a single vendor's platform rather than across vendors. Some resources require an entity tag; others are subject to replication delay that makes an immediate re-read unsound; many offer neither primitive. **A platform-wide assumption is not evidence.** Each target adapter must document and test the exact primitive available for the declared operation, and the assurance mode follows from that test rather than from the vendor's name.

The consequence for this architecture is that a design presupposing compare-and-set at the target would be unimplementable for a large share of tier 4 work. The companion specification says so, and defines a weaker, explicitly declared alternative rather than assuming the primitive exists.

Where a target cannot support a conditional commit, the assurance claim must be lowered. A read-then-write adapter can detect many changes but still contains a race. Compensating transactions can repair some outcomes but do not make an irreversible action unhappen. The evidence record must therefore state the execution mode used.

3.7 Signatures, time and long-term evidence have a lifecycle

A valid signature demonstrates that a holder of a signing key signed a particular representation. It does not by itself prove that the signer's clock was correct, that the key was uncompromised at the relevant time, that the certificate was valid then, or that the cryptographic algorithm will remain trustworthy for the retention period.

Long-lived evidence therefore needs more than a current signature: a defined and monitored time source; a trustworthy timestamp where the timing assertion matters; signing-key separation, rotation and revocation; preservation of certificate chains, revocation information and verification metadata; algorithm agility and re-sealing before an algorithm or key becomes unsuitable; documented handling of key compromise; and independent checkpoints for evidence that must survive compromise of the operating domain.

RFC 3161 provides a time-stamp protocol for proving that a datum existed before a particular time; RFC 5816 updates it to permit `ESSCertIDv2`, which allows signer certificates to be identified with hash algorithms other than SHA-1. RFC 4998 defines Evidence Record Syntax for preserving existence and integrity evidence over long archival periods, with an XML variant in RFC 6283 and alternative ASN.1 modules in RFC 9169. These are building blocks, not a complete deployment design. An implementation must state which time, signature and renewal profile it uses, and must avoid the broad claim of non-repudiation where the operational prerequisites are not met.

3.8 What the control plane does answer

> **How can organisational authority and technical execution be connected so that AI-assisted actions can be controlled, bounded and evidenced?**

It is one component of a broader governance and security architecture. What becomes indispensable is not a product labelled "runtime control plane" — it is the architectural capability to connect identity, mandate, policy, approval, execution and evidence at runtime.

---

4. Governed Intelligence: a working definition

4.1 From governance documents to executable governance

Governance in most organisations is a body of documents: policies, role descriptions, approval processes, risk assessments, training material, procedures, audit reports. These remain necessary. Their effect is bounded when the underlying rules cannot be evaluated or enforced during technical execution.

> **A policy describes what should happen. A control architecture decides whether it may happen. Evidence shows what did happen.**

Governed Intelligence connects the three.

4.2 Definition

> **Governed Intelligence** is an operating state in which an AI-assisted action passes an enforceable, reviewable and revocable control chain of intent, identity, mandate, context, policy, approval, execution and evidence.

The chain:

> **Intent → Identity → Mandate → Context → Policy → Approval → Execution → Evidence**

4.3 The eight elements

1. Intent — what is supposed to happen. The original instruction is captured as a structured intention: not only the prompt, but the intended business transaction. In practice a language model usually performs the structuring, which would put the model in the authority path if left unaddressed. The resolution is explicit: model-assisted structuring produces a **proposal**, never an authority artefact; the proposal is validated against a predefined intent schema, and anything outside the schema is rejected rather than interpreted; for tier 3 and above the structured intent is confirmed by the principal in human-readable form; and the evidence record marks the intent's origin as model-generated, user-confirmed or system-generated.

2. Identity — who or what is acting. Every relevant entity needs a distinct identity: natural person, business role, application, service, agent, agent instance, model, external provider. An agent must not appear as an anonymous process behind a general service account. This is workload identity as understood by SPIFFE/SPIRE, extended to agent instances and their parent-child relationships. Real agent topologies are chains — a planner calls an executor, which calls a sub-agent, which calls a tool server — and an identity model that records only one hop cannot express who acted on whose behalf.

3. Mandate — on whose behalf and with what authority. The mandate links an agent identity to a legitimate principal, and defines the principal, permitted purpose, scope of authority, permitted systems and tools, time bounds, monetary or risk limits, delegation rules and revocation.

> A technical entitlement answers: *can this system reach that?* A mandate answers: *may it take this action, for this principal, for this purpose?*

A per-action ceiling is not by itself a limit: a thousand actions of 999 euro each under a 1,000-euro limit are all individually within the mandate. A mandate needs a cumulative budget and a velocity bound, not only a maximum single value. Mandates should be as short-lived and as narrowly scoped as practical.

4. Context — under what conditions. Tenant and legal entity, business unit and cost centre, data classification, affected data subjects, geographic region, model or provider location, target system, risk class, transaction value, current security posture, and prior decisions in the same process. The same agent may hold different rights in different contexts.

5. Policy — which rules apply now. The policy engine evaluates intent, identity, mandate and context against versioned rules. Possible outcomes: permit; deny; reduce data scope; require a different model; enable additional logging; require human approval; require a second approval; escalate to a specialist. Every decision must be traceable to the specific policy version in force at decision time.

6. Approval — which human authorisation is required. Not every action needs manual approval; the requirement must be risk-based against the tier model. What matters is that the approval is **bound** to the specific action — content, recipient, amount, time and target resource — and, per §3.2, to the **provenance** of each parameter, and, per §3.6, to the material execution preconditions and an expiry.

Two properties of that binding are easy to state and hard to implement, and both are addressed in the companion specification. The approval should bind to *what the human was actually shown*, not merely to a canonical data structure that a compromised interface could render differently — and that is achievable only where the approver's own client does the rendering and the signing. Where a central service both draws the screen and computes the digest, the binding proves what that service committed to rendering, which is a weaker claim and must be stated as one. And for tier 5, a single approver field is not a control at all: quorum, role separation and a rule that the approver is neither the requester nor the beneficiary are what make four-eyes real.

7. Execution — what actually runs. The action is not executed through unbounded, permanently stored credentials, but through controlled execution points: action broker, tool gateway, policy enforcement point, short-lived tokens, transaction-bound credentials, permitted command profiles, narrow API scopes. Immediately before commit, the enforcement point revalidates mandate status, policy freshness, approval expiry and material target-state preconditions.

8. Evidence — what can be proven later. An evidence record links instruction, identities, mandate, context, policy decision, approvals, executed action, parameter provenance, precondition state, result, decision and execution timestamps, cryptographic profile and relevant versions.

Evidence is more than a log. A log shows a technical event. Evidence establishes the relationship between organisational authority and technical action. It also has to survive the party it incriminates: a broker that executes an action and then simply omits the record leaves nothing behind, so the record of intent to commit must be written *before* the commit, and the sequence must be structured so that a missing entry is visible as a gap rather than as silence.

4.4 Twelve control invariants

The eight elements describe the chain. The following invariants describe the minimum properties that must remain true regardless of product choice or deployment style. They are the canonical statement; the *GAIP Candidate Specification* elaborates each into testable requirements and carries a traceability matrix from these twelve to its own rules, its processing steps and its negative tests.

ID	Invariant	Failure that it prevents or exposes
GI-1 Principal attribution	Every governed action resolves to a human or organisational principal and a distinct acting workload, through the full delegation chain	Anonymous service-account action
GI-2 Authority separation	Business applications and models cannot mint, widen or self-approve their own mandates	Self-escalation and circular approval
GI-3 Bypass declaration and detection	Every path to a tier 4 or 5 target operation outside the enforcement point is inventoried, and divergence between target state and evidence is detected by reconciliation	Undeclared side door treated as impossible
GI-4 Proposal is not authority	Model-produced intents and parameters remain proposals until schema validation and required confirmation	Model entering the authority path

ID	Invariant	Failure that it prevents or exposes
GI-5A Approval artefact equivalence	The approval signature binds to the exact canonical action, the derived provenance, the execution constraints and the signed rendering artefact	Approval reused for a different action
GI-5H Human-view equivalence	The artefact the approver's own client rendered and signed is the artefact the approver saw	A compromised display showing one value and signing another
GI-6P Preventive state freshness	The target atomically enforces the approved precondition, or a lock is held through commit, so a stale action **cannot** commit within the declared lock scope	TOCTOU execution under changed state
GI-6D Detective state divergence	Where the target offers no such primitive, the action may commit under changed state, but the divergence is detected and escalated within a declared maximum interval	An undetected stale commit
GI-7 Revocation before commit	Mandate, credential, policy and approval revocation are checked at the last practical point before execution	Action continuing after leaver event or incident
GI-8 Evidence independence	No party in the transaction — including the broker that executes it — can silently omit or rewrite the decision-and-execution history	Self-authored audit trail
GI-9 Tenant and key isolation	A tenant compromise does not confer another tenant's authority, keys or evidence	Cross-tenant blast radius
GI-10 Safe failure	Loss of an authority component moves each impact tier to a pre-declared safe state that is tested	Fail-open behaviour discovered during incident

Five invariants changed since v0.4, two of them by splitting, and the honest ones are concessions. **GI-3** previously read "no high-impact bypass". Against software-as-a-service targets that is not implementable: native user interfaces, administrative consoles, other integrations, batch jobs and the vendor's own support staff cannot be closed off by a customer's enforcement point. Stating an unachievable invariant would have made every honest implementation non-conformant and every conformant claim untrue, so the invariant is now declare-and-detect, and the detective control is the auditable part. **GI-8** previously constrained "the application that benefits from an action", which left out the enforcement point itself — the component best placed to suppress a record. GI-1 now runs through the full delegation chain.

GI-5 and GI-6 were each split in v0.5.2, and the reason is worth stating plainly. Both were single invariants asserting a strong property, while the companion specification permitted profiles that cannot deliver it. A single "approval equivalence" invariant claimed binding to what the approver *saw*, when the common implementation — a rendering service that both draws the screen and computes the digest — can only prove what that service *committed to rendering*; if it is compromised it can display one amount and digest another, and every check still passes. A single "state freshness" invariant claimed that stale approvals are rejected at commit, when the fallback mode for targets that offer no conditional write commits unconditionally and finds divergence afterwards.

Calling those the "detective form" of the same invariant would hide a real loss of assurance. They are different properties, and they are now different invariants. An implementation states which it holds, per action type. For tier 5 — critical or irreversible actions — **GI-5H and GI-6P are the requirement**, because detecting an irreversible stale action after it has executed is evidence, not control.

A system does not need to be a single product to satisfy these invariants. A composed estate can satisfy these invariants if the boundaries and proofs are real. Conversely, a platform marketed as a control plane does not satisfy them merely because it contains components with these names.

5. Regulatory convergence becomes an architecture problem

5.1 No single regulation requires a control plane

Neither the AI Act nor the GDPR, NIS2 or DORA prescribes a Vianordis-like platform. The instruments pursue different objectives. The AI Act addresses risks of specific AI systems and the roles along the AI value chain. The GDPR protects personal data and the rights of data subjects. NIS2 and the German BSI Act require appropriate cybersecurity and risk management from in-scope entities. DORA specifies digital operational resilience for the financial sector. The Data Act governs access to and sharing of data from connected products, and cloud switching. The Cyber Resilience Act governs security properties and vulnerability handling for products with digital elements.

At the technical level these objectives keep meeting at the same points: identities and roles, data access, third parties, technical entitlements, approvals, logging, monitoring, risk decisions and demonstrability.

The convergence does not arise because the laws are alike. It arises because their implementation lands on the same IT systems and business processes. A shared technical control layer can satisfy several of them with fewer separate proofs — which is an efficiency argument, not a legal necessity.

5.2 EU AI Act: what applies when

Regulation (EU) 2024/1689 as amended by Regulation (EU) 2026/1744 (the Digital Omnibus on AI, adopted 8 July 2026, published in the Official Journal on 24 July 2026, in force since 27 July 2026).

5.2.1 Application timeline

Scope	Applies from
Chapters I and II — general provisions, prohibited practices (Art. 5), AI literacy (Art. 4)	2 February 2025
Chapter III Section 4, Chapter V (GPAI models), Chapter VII (governance) and Chapter XII (penalties, except Art. 101), Art. 78	2 August 2025
GPAI models placed on the market before 2 August 2025 — compliance deadline under Art. 111(3)	2 August 2027
General applicability; Art. 50 transparency; Chapter VI; Art. 101 GPAI fines	**2 August 2026**
Transitional deadline for Art. 50(2) machine-readable marking, for synthetic-content systems already placed on the market before 2 August 2026	2 December 2026
New Art. 5 prohibitions introduced by the Digital Omnibus (non-consensual intimate imagery, CSAM)	2 December 2026
Chapter III Sections 1–3 — the core high-risk obligations — for Art. 6(2) / Annex III systems	**2 December 2027**
Chapter III Sections 1–3 for Art. 6(1) / Annex I systems (safety components in regulated products)	**2 August 2028**
AI regulatory sandboxes (Art. 57) — Member State obligation to establish, deferred by one year; not an application date for obligations	2 August 2027

The deferral is unconditional. The Commission's November 2025 proposal had tied the dates to a finding on the availability of harmonised standards; that trigger is not in the final Article 113.

Two consequences for planning. First, the obligations most often cited as the reason to act now — logging, human oversight, deployer duties — do not apply until December 2027 at the earliest. Second, that is roughly one budget cycle plus one implementation cycle, which for an estate-wide architecture change is not generous.

5.2.2 What the high-risk regime will require

For systems in scope from 2027/2028, the AI Act provides among other things:

- **Art. 12** — the system must technically enable automatic recording of events over its lifetime. Art. 12 is a design obligation; retention follows from Art. 19 and Art. 26(6).
- **Art. 14** — effective human oversight must be *enabled by the provider*. Under Art. 14(4), oversight persons must be able to correctly interpret the output, "to decide, in any particular situation, not to use the high-risk AI system or to otherwise disregard, override or reverse the output", and "to intervene or interrupt the system through a 'stop' button or a similar procedure". Art. 14(5) adds a four-eyes requirement for remote biometric identification.
- **Art. 19** — providers retain automatically generated logs under their control for **at least six months**, with a special rule for financial institutions.
- **Art. 26** — deployer obligations: use in accordance with the instructions, appropriate technical and organisational measures, assignment of human oversight to natural persons "who have the necessary competence, training and authority" (Art. 26(2)), monitoring of operation and reporting (Art. 26(5)), and log retention of at least six months (Art. 26(6)).
- **Art. 25** — role reallocation. A party that puts its name or trademark on a high-risk system, makes a substantial modification, or changes the intended purpose such that the system becomes high-risk, becomes the provider of the resulting system. Art. 25(4) contains a separate carve-out for tools and components supplied under a free and open-source licence.
- **Art. 27** — the fundamental rights impact assessment is retained, and may be satisfied by cross-reference to a DPIA to the extent the Art. 27 requirements are covered there.

Note the division of labour in Art. 14 and Art. 26: **enabling** oversight is a provider obligation; **assigning** it to competent people is a deployer obligation. Organisations that integrate, rebrand or substantially modify a system may find they hold both.

5.2.3 Changes introduced by the Digital Omnibus that are easy to miss

- **Art. 4 (AI literacy) weakened** — from "shall ensure … a sufficient level of AI literacy" to an obligation to take measures to support the development of AI literacy. An obligation of means, not of result.
- **"Safety component" narrowed** (Art. 3(14) and Art. 6) — systems that concern only user assistance, performance optimisation, efficiency, automation, comfort or quality control are not safety components. This materially reduces the Annex I high-risk perimeter.
- **New Art. 5 prohibitions from 2 December 2026** — AI systems for generating or manipulating non-consensual intimate imagery and child sexual abuse material, including where this is "reasonably foreseeable" without substantial technical modification. Penalty band: 35 million euro or 7 %. These prohibitions require effective technical and organisational safeguards; documentation alone is insufficient, and runtime controls may be necessary depending on the system's capabilities and its reasonably foreseeable use. Product restrictions, model capability limits, distribution and account controls, moderation, detection and reporting all form part of the available answer.
- **New Art. 4a and adjusted Art. 10(5)** — processing of special categories of personal data for bias detection and correction, extended to providers and deployers of non-high-risk systems and models, subject to safeguards including pseudonymisation, access controls and deletion. A direct GDPR interface.
- **Art. 75 — enforcement shifts to the AI Office** for AI systems built on a GPAI model of the same provider and for systems integrated into VLOPs/VLOSEs, with extended investigation and fining powers.
- **Registration retained.** The proposal to remove the Art. 6(3) / Art. 49 registration duty for systems self-assessed as non-high-risk was **not** adopted; the duty remains, with Annex VIII Section B items 7 and 9 removed.
- **New "small mid-cap" category** — simplified quality management and technical documentation, priority sandbox access, and capped penalties.
- **Machinery Regulation (EU) 2023/1230 moved** from Annex I Section A to Section B, with a delegated act due by 2 August 2028.

5.2.4 Penalties

Band	Amount	Scope
Art. 99(3)	up to **35 million euro or 7 %** of total worldwide annual turnover, **whichever is higher**	Non-compliance with the Art. 5 prohibitions
Art. 99(4)	up to **15 million euro or 3 %**, whichever is higher	Most other obligations of providers, deployers, importers, distributors and notified bodies
Art. 99(5)	up to **7.5 million euro or 1 %**, whichever is higher	Supplying incorrect, incomplete or misleading information to notified bodies or national authorities
Art. 99(6)	for SMEs and start-ups, each band is capped at the **lower** of the two figures	Applies across paragraphs 3, 4 and 5
Art. 99(6a)	for **small mid-cap enterprises**, each fine under **Art. 99(4) and Art. 99(5)** is capped at the lower of the applicable fixed amount and turnover percentage	**Does not extend the lower cap to the Art. 99(3) prohibition band**
Art. 101	up to **3 % of annual total worldwide turnover or 15 million euro, whichever is higher**	Commission fines on GPAI model providers; applicable from 2 August 2026

The Digital Omnibus left the amounts and the tier structure unchanged, and adjusted which infringements fall into the Art. 99(4) band, pulling breaches of Art. 25(2) and 25(4) into it. It inserted the small mid-cap cap as a new Art. 99(6a). Note the asymmetry with the SME rule: the SME cap in Art. 99(6) runs across paragraphs 3, 4 and 5, while the small mid-cap cap in Art. 99(6a) reaches only paragraphs 4 and 5. **A small mid-cap that breaches an Article 5 prohibition faces the full 35 million euro or 7 % band.**

5.2.5 Open source under the AI Act

Art. 2(12) excludes AI systems released under free and open-source licences from the Regulation's scope, unless they are placed on the market or put into service as high-risk systems, or as systems falling under Art. 5 or Art. 50. Three limits matter for any open-source control plane, including ours:

1. The exclusion covers **AI systems**, not **GPAI models**. Models fall under the narrower reliefs of Art. 53(2) and Art. 54(6), which fall away entirely where systemic risk is present.
2. Monetised or not-genuinely-open distributions are not covered, per the recitals. Whether a dual-licensing construction still falls within Art. 2(12) is a real question and not one we consider settled.
3. Art. 25(4) provides a separate FOSS carve-out for value-chain suppliers.

5.3 Germany: NIS2 and the BSI Act — management liability is already live

The NIS2 implementation act of 2 December 2025 (BGBl. 2025 I no. 301, issued 5 December 2025) brought the new BSI Act (BSIG 2025) into force on **6 December 2025**, with no general transition period for the substantive duties.

§ 30 BSIG requires essential and important entities to take appropriate, proportionate and effective technical and organisational risk management measures, proportionate to risk exposure, size, cost and potential damage, and to **document** them. § 30(2) lists ten areas: risk analysis and security concepts; incident handling; business continuity, backup and crisis management; supply chain security; security in acquisition, development and maintenance of systems; effectiveness assessment; training and cyber hygiene; cryptography and encryption; personnel security, access control and asset management; and multi-factor authentication, secured communications and emergency communications.

§ 38 BSIG is the provision that changes the conversation with a board:

- **§ 38(1)** — management bodies of essential and important entities are obliged to implement the § 30 risk management measures and to supervise their implementation.
- **§ 38(2)** — management bodies that breach those duties are liable to their entity for culpably caused damage, under the company-law rules applicable to the entity's legal form.

- **§ 38(3)** — regular training obligation for the management body itself.

This is internal liability toward the entity, fault-based, measured by the standard applicable to the legal form (§ 43 GmbHG, § 93 AktG and equivalents). It is distinct from administrative fines against the entity under § 65 BSIG. It is in force now, unlike the AI Act's high-risk regime.

Staged deadlines that follow from the Act and are easy to miss:

Duty	Deadline
Registration (§ 33 BSIG)	within **three months** of first becoming in scope — for entities in scope at entry into force, generally **6 March 2026**
BSI portal for registration and reporting	live since 6 January 2026
Notification of changes	without undue delay, at the latest within two weeks
Incident reporting (§ 32)	**24 h** early warning / **72 h** report / **one month** final report
KRITIS evidence (§ 39(1))	first at a date set by the BSI, thereafter every three years; § 39(3) contains a transition rule for entities designated under the 2009 BSIG

The 24-hour early-warning clock is itself an argument for correlated evidence. In an agent-driven incident, the first 24 hours are spent determining what the agent did, on whose behalf and with what data. Organisations that have to join four log sources by hand to answer that will miss the window.

A further amendment is pending: the draft CRA implementation act (BT-Drs. 21/6134 of 26 May 2026) designates the BSI as market surveillance and notifying authority.

5.4 DORA: responsibility survives outsourcing

Regulation (EU) 2022/2554 has applied to in-scope financial entities since **17 January 2025**.

Art. 5(1) requires an internal governance and control framework ensuring effective and prudent management of ICT risk. Art. 5(2) provides that the management body "shall define, approve, oversee and be responsible for the implementation of all arrangements related to the ICT risk management framework", and Art. 5(2)(a) assigns it **ultimate responsibility**. Art. 6(1) requires a "sound, comprehensive and well-documented ICT risk management framework"; Art. 6(4) requires an independent control function with separation along three-lines principles. Art. 16 disapplies Art. 5 to 15 for a closed list of entities — small and non-interconnected investment firms, certain exempted payment, credit and e-money institutions, and small IORPs — and substitutes a simplified framework. Microenterprise reliefs sit elsewhere, in Art. 5(2), 6(4) and 28(2). The general statement therefore does not apply uniformly to all financial entities.

On third parties, Art. 28(1)(a) is the operative sentence: financial entities "shall at all times remain fully responsible for compliance with, and the discharge of, all obligations under this Regulation and applicable financial services law". Art. 28(1) requires integration of ICT third-party risk into the Art. 6(1) framework specifically; Art. 28(2) requires a documented strategy reviewed regularly by the management body; Art. 28(3) requires the register of information. Art. 28(8) governs **exit strategies** for ICT services supporting critical or important functions; **concentration risk** is the subject of Art. 29, with Art. 28(4)(c) treating it as an assessment input.

Applied to AI:

> **Procuring a model, a cloud platform or an agent service does not transfer enterprise responsibility to the supplier.**

An organisation therefore needs more than a list of its AI suppliers. It needs a controllable view of the actions, data flows, dependencies and entitlements those suppliers enable.

5.5 GDPR: purpose and minimisation have to hold at execution time

The GDPR applies unchanged. Note for completeness: the Digital Omnibus proposal on data and privacy of 19 November 2025 — a **different instrument** from the AI Omnibus above — remains in the legislative process as at August 2026. Proposed changes to pseudonymisation, legitimate interest as a basis for AI training, records of processing and breach notification timelines are **not** current law and should not be planned against as if they were.

Art. 25(1) requires the controller to implement appropriate technical and organisational measures both "at the time of the determination of the means for processing" and "at the time of the processing itself" — narrower than a continuous-supervision duty, and worth quoting precisely rather than paraphrasing.

Art. 25(2) — data protection by default — is the more directly applicable provision for agents, and is frequently overlooked. It requires that, by default, only personal data necessary for each specific purpose are processed, in terms of amount collected, extent of processing, storage period and accessibility. For an agent architecture this reads almost as a specification: default entitlements, default retention, default logging scope and default data breadth all fall under it.

Alongside these, records of processing activities under Art. 30 may be required. For organisations employing fewer than 250 persons, the exemption in Art. 30(5) is unavailable where the processing is not occasional, is likely to result in a risk to the rights and freedoms of data subjects, or includes data under Art. 9 or Art. 10. **AI use alone does not remove the exemption**, although many persistent or sensitive enterprise AI processes will fall within one or more of those exceptions. A data protection impact assessment may be required depending on the Art. 35 criteria and the applicable supervisory authority's lists, typically via Art. 35(3)(a) where systematic and extensive evaluation or profiling is present.

An agent with technical access to a large corpus should therefore not be permitted to use all of it for every task. The control architecture must be able to take purpose, context and data classification into account at decision time.

5.6 Two instruments often left out

5.6.1 Cyber Resilience Act — Regulation (EU) 2024/2847

The CRA applies to "products with digital elements" placed on the EU market. Scope is narrower than often assumed: **standalone software-as-a-service is generally outside it** unless it constitutes an integral remote data processing solution — the operative test is in Art. 3(2), which catches remote data processing "the absence of which would prevent the product from performing one of its functions". What is otherwise caught is distributable or on-premises software.

Being outside the CRA does not place a provider inside NIS2 automatically. A cloud or SaaS provider falls within NIS2 only where the entity, sector and size criteria of that Directive are met. The two instruments have different scoping tests, and a provider can fall under one, both or neither.

The date that matters immediately: from **11 September 2026**, manufacturers must report **actively exploited vulnerabilities** under Art. 14(1)–(2) on a 24-hour early-warning / 72-hour notification / final-report-no-later-than-14-days-after-a-corrective-or-mitigating-measure-is-available sequence. **Severe incidents** follow Art. 14(3)–(4): 24 hours, 72 hours, and a final report within one month of the incident notification. The notification framework — Chapter IV, Art. 35–51 — has applied since 11 June 2026, and the Regulation applies in full from **11 December 2027**.

Two provisions are frequently conflated and should not be. **Art. 14 is the reporting obligation**, applicable from 11 September 2026. The **vulnerability-handling requirements** — coordinated disclosure, SBOM, security updates, remediation — are in **Annex I Part II**, made binding through **Art. 13(8)**, and apply only from **11 December 2027**. (Art. 13(6) is a different duty: reporting a vulnerability found in an integrated third-party or open-source component back to that component's maintainer.) Reporting comes first by fifteen months.

For this paper's argument the CRA matters twice over: it imposes on suppliers of installable control-plane software exactly the supply-chain and vulnerability-handling discipline this architecture presumes, and it gives enterprise buyers a lawful basis to demand it.

5.6.2 Data Act — Regulation (EU) 2023/2854

Applicable since 12 September 2025, with design obligations for new connected products from 12 September 2026 and cloud-switching provisions — including the removal of switching charges — from 12 January 2027. This is the instrument that turns provider portability from a preference into a legal entitlement. An organisation arguing internally for provider-independent identities, policies and evidence now has a regulation behind it rather than only an architectural principle.

5.6.3 Worth watching: eIDAS 2.0

Regulation (EU) 2024/1183 requires Member States to make EU Digital Identity Wallets available by end-2026, with acceptance obligations for certain private service providers from late 2027 (36 months from entry into force of the core implementing acts). For the "who is actually acting" question in §4.3, this is the most relevant identity development in the EU and should be tracked by anyone designing an agent identity model now.

5.7 Shared technical objectives

Regulatory objective	Possible technical support
Accountability	Distinct identities for people, services and agents
Access control	Short-lived, purpose-bound entitlements
Purpose limitation	Mandates with a stated purpose and permitted context
Human oversight	Risk-based approvals, escalation and interruption
Traceability	Correlated event and decision records
Resilience	Controlled execution, repeatability, blocking and restart mechanisms
Third-party governance	Model and provider abstraction with documented data paths
Demonstrability	Versioned policies, bound approvals, evidence records
Data minimisation	Context-dependent limitation of the information supplied
Supply chain security	Traceable components, dependencies and signed releases
Vulnerability handling	Coordinated disclosure, SBOM, reporting readiness within 24 h

This mapping does not create compliance. It shows where a shared architecture can provide reusable controls. Annex A maps the same capabilities against individual instruments.

6. Conclusion

The next phase of artificial intelligence will not be determined by more capable models alone. It will be determined by how reliably organisations draw the line between **proposal and action**.

An organisation can use an excellent model and still lose control when identities are unclear, mandates are absent, entitlements are too broad, approvals are unbound, policies exist only on paper, business applications can bypass central controls, and logs do not form a shared, tamper-evident chain.

The central question is therefore not *"how autonomous can our AI become?"* but:

> **"What autonomy can we take responsibility for — controllably, revocably and demonstrably?"**

Governed Intelligence describes an architecture in which AI does not act merely because it is technically capable of acting. It acts inside a reviewable authority structure, along the control chain set out in §4.2.

Whether that capability is delivered by a single platform or by several tightly integrated components is an architecture decision, and in many estates assembling it from existing standards is the right answer. What is not optional is the end-to-end connection of identity, mandate, context, policy, human authorisation, controlled execution and defensible evidence — with parameter provenance, because without it the evidence answers the wrong question.

Vianordis is our attempt at an open European reference implementation of this architecture class. The release-status page states what exists today, which is less than this paper describes.

The thesis of this paper is deliberately testable:

> **Enterprises do not fail to scale AI only because of model quality or regulation. They fail because organisational authority and technical execution are not connected.**

The *Adoption & Measurement Guide* states four measurements that would confirm, weaken or falsify material parts of it. We do not have them yet. Developing a defensible answer should not happen behind closed doors.

Vianordis therefore invites enterprises, IT leaders, security officers, data protection professionals, lawyers, developers, researchers and the open-source community to examine, criticise and develop both the proposed architecture and the GAIP candidate profile — through the contact route below. Substantive critique will be reflected in the next edition with attribution, unless the contributor prefers otherwise.

Annex A — Control capabilities and regulatory objectives

This matrix is orientation, not a conformity assessment. Article references are indicative and, for AI Act high-risk provisions, apply from 2 December 2027 or 2 August 2028 as set out in §5.2.1. NIST AI RMF identifiers below are **categories**; subcategories carry a decimal place (GOVERN 1.1, MAP 4.1, MEASURE 2.5) and are not cited here. ISO/IEC 27001 references are to the 2022 edition as amended by Amd 1:2024.

Control capability	EU AI Act	GDPR	NIS2 / BSIG 2025	DORA	Management-system standards
Distinct identities	Provider and deployer responsibilities, Art. 16, 26	Accountability, Art. 5(2); processing under instruction, Art. 28	Access control, § 30(2) no. 9	Clear roles, Art. 5(2)	ISO/IEC 27001 A.5.15–A.5.18; ISO/IEC 42001 cl. 5
Mandates and least privilege	Controlled use and human oversight, Art. 14, 26	Purpose limitation and minimisation, Art. 5(1)(b),(c); by default, Art. 25(2)	Appropriate technical and organisational measures, § 30(1)	ICT risk management and access protection, Art. 9(4)(c),(d)	ISO/IEC 27001 A.5.15; NIST AI RMF GOVERN 1 (category)
Automatic logging	Art. 12 (design obligation); retention Art. 19, 26(6)	Accountability and documentation, Art. 5(2), 30	Documentation and evidence duties, § 30(1)	Documented ICT risk framework, Art. 6(1)	ISO/IEC 27001 A.8.15; NIST AI RMF MEASURE 2 (category)
Human approval	Art. 14, 26(2)	Safeguards per processing context; Art. 22 where applicable	Management responsibility, § 38(1)	Management body responsibility, Art. 5(2)	ISO/IEC 42001 cl. 8; NIST AI RMF MANAGE 4 (category)
Parameter provenance	Supports Art. 12 record quality and Art. 14 interpretability	Supports accuracy, Art. 5(1)(d)	Supports incident analysis, § 32	Supports incident reporting, Art. 19	NIST AI RMF MAP 4 (category)
Precondition and freshness binding	Supports reliable controlled use and oversight	Supports accuracy and purpose limitation at processing time	Supports effectiveness and incident containment	Supports operational resilience and controlled change	ISO/IEC 27001 change and access controls
Provider abstraction	Roles along the AI value chain, Art. 25	Processors and third-country transfers, Art. 28, 44 ff.	Supply chain security, § 30(2) no. 4	ICT third-party risk, Art. 28	ISO/IEC 27001 A.5.19–A.5.22
Evidence record	**Traceability — Art. 12, 19, 26(6); cooperation with authorities — Art. 21**; post-market monitoring, Art. 72	Demonstrability, Art. 5(2); DPIA, Art. 35	Documentation, audit, incident handling, §§ 30, 32	Governance, reporting and audit, Art. 5, 6	ISO/IEC 27001 A.5.28; ISO/IEC 42001 cl. 9
Cryptographic tamper-evidence	Supports integrity of technical evidence	Supports accountability; constrained by minimisation and erasure	Cryptography, § 30(2) no. 8	Supports controllable audit trails	ISO/IEC 27001 A.8.24
Trusted time and key lifecycle	Supports reliable timing and retained verification of logs	Supports demonstrability; retention must remain proportionate	Supports incident timelines and cryptographic governance	Supports auditability and resilience	RFC 3161 / RFC 5816; RFC 4998 / RFC 6283; ISO/IEC 27001 cryptographic controls
Lifecycle and revocation	Monitoring, correction and oversight, Art. 26(5), 72	Storage limitation, Art. 5(1)(e); change of risk, Art. 35(11)	Effectiveness review, § 30(2) no. 6	Continuous monitoring and improvement, Art. 6(5)	ISO/IEC 42001 cl. 10

Control capability	EU AI Act	GDPR	NIS2 / BSIG 2025	DORA	Management-system standards
Model and provider substitution	Avoids unclear role and integration dependencies, Art. 25	Controllable data paths	Supply chain and dependency management, § 30(2) no. 4	**Exit strategies — Art. 28(8); concentration risk — Art. 29** (Art. 28(4)(c) as assessment input)	— (Data Act Art. 23–31 also applies)
Core isolation	Supports controlled technical use	Supports privacy by design, Art. 25(1)	Segmentation, access control, secure administration, § 30(2) nos. 5, 9	Supports ICT risk containment and resilience, Art. 9	ISO/IEC 27001 A.8.20–A.8.23
Vulnerability handling	—	Art. 32, 33 where personal data affected	§ 32 reporting	Art. 17–19 incident management	**CRA Annex I Part II via Art. 13(8), applicable 11 December 2027; reporting CRA Art. 14, applicable 11 September 2026**; ISO/IEC 27001 A.8.8

Annex B — Claims this architecture must not make

A defensible architecture states what it trusts and what it cannot prove. The full trust-boundary analysis and the compromise scenarios this design aims to bound are in the *GAIP Candidate Specification*. What belongs in the position paper is the shorter list: the shorthand claims that exceed what the controls actually establish.

Prohibited shorthand	Defensible statement
"Immutable audit log"	Subsequent alteration, gaps or reordering are detectable under the stated trust assumptions
"The approval proves the payment was correct"	The approval proves that the approver accepted the displayed action and provenance under the recorded preconditions
"Signed means trusted"	The object verifies under a particular key and validation profile; key trust and compromise status still matter
"The AI was authorised"	A specific agent instance was authorised for a bounded action on behalf of a principal
"EU hosted means sovereign"	The deployment satisfies the separately stated controls over identities, keys, data paths, providers and exit
"Open source means compliant"	Source can be inspected; compliance depends on system, deployment, organisation and use
"No bypass"	Alternative paths to the declared action set are inventoried, and divergence is detected by reconciliation; the claim has a scope and a date
"GAIP conformant"	Not currently a meaningful statement. The candidate specification defines no conformance levels and publishes no certification scheme

These distinctions are not semantic caution. They determine what an auditor, customer or incident investigator can rely on. The last row is new in this edition: v0.4 proposed five named conformance levels, and they have been withdrawn for the reasons set out in the companion specification.

The three compromise cases that most deserve a board's attention, because the architecture bounds rather than prevents them:

- **Input manipulation.** Not prevented. A model may derive falsified parameters and a correctly bound approval will bind to them (§3.2).
- **A fully compromised administrative domain.** Bounded by key separation, four-eyes controls and independent evidence checkpoints — not eliminated.
- **State change after approval.** The approval becomes stale; whether that is detected before commit depends on primitives the target system may not offer (§3.6).

Annex C — Primary sources

Legal and regulatory instruments cited in this paper, with links to the official texts. The standards and RFC list — policy languages, delegation formats, canonicalisation, signature containers, transparency logs and supply-chain attestation — is in the *GAIP Candidate Specification*, where those references do the work.

European Union

1. Regulation (EU) 2024/1689 — Artificial Intelligence Act: https://eur-lex.europa.eu/eli/reg/2024/1689/oj
2. Regulation (EU) 2026/1744 — Digital Omnibus on AI, amending Regulation (EU) 2024/1689; adopted 8 July 2026, OJ 24 July 2026, in force 27 July 2026: https://eur-lex.europa.eu/eli/reg/2026/1744/oj · consolidated AI Act as amended: https://eur-lex.europa.eu/eli/reg/2024/1689/2026-07-27/eng
3. Regulation (EU) 2016/679 — General Data Protection Regulation: https://eur-lex.europa.eu/eli/reg/2016/679/oj
4. Directive (EU) 2022/2555 — NIS2: https://eur-lex.europa.eu/eli/dir/2022/2555/oj
5. Regulation (EU) 2022/2554 — DORA: https://eur-lex.europa.eu/eli/reg/2022/2554/oj
6. Regulation (EU) 2024/2847 — Cyber Resilience Act: https://eur-lex.europa.eu/eli/reg/2024/2847/oj
7. Regulation (EU) 2023/2854 — Data Act: https://eur-lex.europa.eu/eli/reg/2023/2854/oj
8. Regulation (EU) 2024/1183 — eIDAS 2.0 / European Digital Identity Framework: https://eur-lex.europa.eu/eli/reg/2024/1183/oj 8a. Commission Delegated Regulation (EU) 2018/389 — PSD2 regulatory technical standards on strong customer authentication; **Art. 5, dynamic linking**: https://eur-lex.europa.eu/eli/reg_del/2018/389/oj
9. Regulation (EU) 2023/1230 — Machinery Regulation: https://eur-lex.europa.eu/eli/reg/2023/1230/oj

Germany

10. BSIG 2025, in force 6 December 2025, via the NIS2 implementation act of 2 December 2025 (BGBl. 2025 I no. 301): https://www.gesetze-im-internet.de/bsig_2025/
11. § 87(1) no. 6 Betriebsverfassungsgesetz — co-determination on technical monitoring devices: https://www.gesetze-im-internet.de/betrvg/__87.html
12. § 43 GmbHG and § 93 AktG — the directors' duty-of-care standards referenced by § 38(2) BSIG: https://www.gesetze-im-internet.de/gmbhg/__43.html · https://www.gesetze-im-internet.de/aktg/__93.html
13. Draft CRA implementation act, BT-Drs. 21/6134 of 26 May 2026

Standards and technical references cited in this document

14. ISO/IEC 42001:2023 — AI management systems; ISO/IEC 27001:2022 as amended by Amd 1:2024 — information security management
15. NIST AI RMF 1.0 (NIST AI 100-1, January 2023): https://doi.org/10.6028/NIST.AI.100-1
16. RFC 3161 — Time-Stamp Protocol; RFC 5816 — ESSCertIDv2 Update for RFC 3161
17. RFC 4998 — Evidence Record Syntax; RFC 6283 — Extensible Markup Language Evidence Record Syntax (XMLERS); RFC 9169 — New ASN.1 Modules for ERS (Informational; supplies alternative modules and does not formally update RFC 4998)
18. RFC 8693 — OAuth 2.0 Token Exchange
19. Hardy, N. (1988). "The Confused Deputy: (or why capabilities might have been invented)". *ACM SIGOPS Operating Systems Review* 22(4), pp. 36–38

Pending legislation — not current law

20. European Commission, Digital Omnibus proposal on data, privacy and ePrivacy, 19 November 2025 — in the legislative process as at August 2026

Annex D — Version history

Version	Date	Language	Principal change	Author
0.1	2026	DE	Initial draft	Y. F. N'Soussoula
0.2	2026-08-14	DE	Institutional stewardship; Core as security boundary; cryptographic tamper-evidence; extended trust-boundary analysis; regulatory sources updated for Regulation (EU) 2026/1744	Y. F. N'Soussoula
0.3	2026-08-14	EN	Prior art, parameter provenance, input manipulation, implementation transparency, maturity model, corrected regulatory timelines	Y. F. N'Soussoula
0.4	2026-08-14	EN	Architecture Profile Edition: Governed Action Envelope; TOCTOU binding; trusted time and key lifecycle; brownfield modes; GAIP-0.1 and measurement protocol	Y. F. N'Soussoula
0.5	2026-08-26	EN	Split into three documents; conformance levels withdrawn; ten citation and attribution corrections; GI-3 restated as declare-and-detect and GI-8 extended to the broker; subtitle changed	Y. F. N'Soussoula
0.5.1	2026-08-26	EN	First external validation pass: board question aligned with GI-3; GDPR Art. 30(5) and CRA/NIS2 scope narrowed; named-platform capability claims withdrawn; novelty claim narrowed against RFC 9396 and PSD2 dynamic linking; works-council conclusion made conditional	Y. F. N'Soussoula
0.5.2	2026-08-26	EN	Second validation pass: Art. 101 corrected; the Art. 99(6a) row withdrawn as unverifiable; GI-5 split into GI-5A/GI-5H and GI-6 into GI-6P/GI-6D so that no headline invariant asserts a guarantee the permitted profiles cannot deliver	Y. F. N'Soussoula
0.5.3	2026-08-26	EN	Third validation pass: Art. 99(6a) restored and correctly scoped to the Art. 99(4) and 99(5) bands; the board brief and executive summary conditioned on GI-6P versus GI-6D; provenance assessment made a status rather than a sixth class	Y. F. N'Soussoula
0.5.4	**2026-08-26**	**EN**	**Fourth validation pass: the unevidenced market-wide claim about enterprise platforms replaced by a test-it requirement; companion references synchronised to GAIP 0.4 and Guide 0.1.4**	**Y. F. N'Soussoula**

What changed in v0.5, in full.

Structure. The paper was divided into three independently versioned documents plus a rolling status page, for the reason given in "How to read this set". The architecture chapter, the prior-art chapter, the trust-boundary annex, the glossary and Annex G moved to the *GAIP Candidate Specification 0.2*. The worked example, adoption path, brownfield modes, maturity model and measurement protocol moved to the *Adoption & Measurement Guide 0.1*. The implementation-status annex moved to the release-status page. The chapter on open source and institutional stewardship is held for separate publication.

Corrections of fact and attribution.

1. **CRA.** Vulnerability handling is **Annex I Part II via Art. 13(8)**, applicable 11 December 2027; **Art. 14** is the reporting obligation, applicable 11 September 2026. v0.4's Annex A attributed vulnerability handling to Art. 14. Corrected in Annex A and §5.6.1, with the fifteen-month sequence now stated explicitly. *An earlier draft of this correction cited Art. 13(6); the operative hook is Art. 13(8), and Art. 13(6) is the separate duty to report a vulnerability found in an integrated component back to its maintainer.*
2. **PDP/PEP attribution.** The vocabulary originates in the IETF policy framework work (RFC 2753, RFC 3198), with antecedents in ISO/IEC 10181-3; XACML adopts it and contributes PIP and PAP. v0.4 credited XACML with PDP/PEP. Corrected in the executive summary and, at length, in the companion specification.
3. **DORA.** Exit strategies are Art. 28(8); **concentration risk is Art. 29**, with Art. 28(4)(c) as an assessment input.
4. **AI Act traceability.** Art. 12, 19 and 26(6), not Art. 21; Art. 21 governs cooperation with authorities.
5. **NIST AI RMF.** GOVERN 1, MAP 4, MEASURE 2 and MANAGE 4 are marked as categories, not subcategories, and the framework is dated (NIST AI 100-1, January 2023).

6. **RFC precision.** RFC 5816 is an ESSCertIDv2 update, not a general TSP revision; RFC 9169 supplies alternative ASN.1 modules and does not formally update RFC 4998; RFC 6283 added.
7. **Conformance levels withdrawn.** GAIP-C/S/E/H/X are removed. A five-level scheme with no test suite, no attestation format, no assessor role and no published test vectors invites a self-certification nobody can refute — and the profile's own rule required test vectors before any interoperability claim.
8. **Board brief.** The version-history paragraph ("What v0.4 adds") was deleted, and the "three obligations … and one" counting error corrected to "four obligations, three of them already in force", with the CRA moved to fourth position so the sentence and the list agree.
9. **Subtitle.** "Why model catalogues, policies and API gateways stop short once AI acts inside the enterprise" became "Why approving a model is not approving an action" — because §1.3 credits those mechanisms rather than dismissing them, and the old subtitle promised a reckoning the paper deliberately does not deliver.
10. **Invariants.** GI-3 restated from "no high-impact bypass" to declare-and-detect, because the absolute form is not implementable against software-as-a-service targets. GI-8 extended from "the application that benefits" to any party in the transaction, including the broker. GI-1 extended to cover delegation chains, and GI-5 extended to cover the rendered presentation — the latter split again in v0.5.2, see below.

Additions. §2.4 and §2.5 now state that tiers and provenance classes must not be self-declared by the requesting component. §3.6 stated which named enterprise systems do not offer a conditional-write primitive; those named claims were withdrawn in v0.5.1, see item 7 below. §4.3 notes the cumulative-budget, four-eyes, delegation-chain and write-ahead-evidence requirements that v0.4 omitted. Annex B gains the "GAIP conformant" row.

What changed in v0.5.1. An external validation review of the v0.5 set produced ten corrections to this document, all of which narrow a claim rather than widen one:

1. **The board brief contradicted GI-3.** Question 3 asked whether bypass was "technically impossible", two chapters before §4.4 explains that the absolute form is not achievable against software-as-a-service targets. The question now asks for an inventory of alternative paths and a reconciliation interval.
2. **GDPR Art. 30(5) was overstated.** v0.5 said the sub-250-employee exemption is "defeated in practice by its own exceptions where AI processing is involved". AI use alone does not remove the exemption; the exceptions in Art. 30(5) — processing that is not occasional, that is likely to risk rights and freedoms, or that includes Art. 9 or Art. 10 data — do. The DPIA sentence was softened from "will frequently be triggered" to a statement of the applicable criteria.
3. **CRA and NIS2 scope.** v0.5 said pure SaaS is outside the CRA "and is addressed by NIS2 instead". A SaaS provider is not automatically in NIS2; that Directive's entity, sector and size criteria still have to be met. The two tests are now stated separately.
4. **Art. 99(6a)** is now quoted in the sourced, band-agnostic form, with an explicit note on the limits of the sourcing.
5. **The Art. 5 prohibitions.** "Can only be met by runtime safeguards" was too absolute; product restrictions, capability limits, distribution and account controls, moderation and detection are also part of the answer.
6. **The works-council conclusion** is now conditional on the concrete fields, attribution model, query functions and deployment, rather than asserted for any implementation of the pattern. "Durable and non-repudiable" became "durable, attributable and cryptographically tamper-evident", because the paper elsewhere warns against exactly the broad non-repudiation claim the old wording implied.
7. **Named-platform claims withdrawn** from §3.6. Conditional-write and idempotency support varies by API, resource and deployment model *within* a vendor's platform, so a platform-wide statement is not evidence for an adapter assurance mode. The requirement is now that each adapter document and test the exact primitive for the declared operation.
8. **The novelty claim is narrower.** Two mechanisms were missing from the prior-art discussion and are now named in the executive summary: **RFC 9396** Rich Authorization Requests, which carries transaction-specific authorisation detail including amount, currency and creditor; and **PSD2 dynamic linking** (Commission Delegated Regulation (EU) 2018/389, Art. 5(1)), which has required since 2019 that a payment authentication code be specific to amount and payee and be invalidated when either changes. Binding an authorisation to a concrete transaction is settled European payments law, not a contribution of this work. Recent agent-governance research is discussed in the companion specification.

9. **"Conform" became "satisfy"** in §4.4, because Annex B states that "GAIP conformant" is not currently a meaningful claim.
10. **A release-status statement** was added to the document-status section.

What changed in v0.5.2. A second external validation pass produced four corrections to this document.

1. **Article 101** now carries "whichever is higher". Its omission changed the legal meaning of the row.
2. **Article 99(6a) has been withdrawn from the penalties table.** Two reviews described the small mid-cap cap differently — one band-agnostic, one restricted to the Art. 99(4) and 99(5) bands. Every reproduction of Article 99 we could reach, including the European Commission's own AI Act Service Desk, still showed only the SME rule in paragraph 6, with the lower cap applying across paragraphs 3, 4 and 5. Rather than print one of two unverified readings, the table now omits the row and the surrounding text says what is known, what is disputed, and that the consolidated Official Journal text governs. The v0.5.1 note claiming the exclusion "could not be confirmed" is replaced by that fuller statement.
3. **GI-5 was split into GI-5A and GI-5H, and GI-6 into GI-6P and GI-6D** (§4.4). The single invariants asserted properties the permitted implementation profiles cannot deliver — human-view equivalence where a central service renders and digests, and preventive freshness where the target offers no conditional write. Tier 5 requires GI-5H and GI-6P.
4. The executive summary's contribution claim and §4.3's approval element were softened to match: binding to a signed rendering artefact in general, and to what the approver saw only where the approver's own client renders and signs.

The underlying issue in items 3 and 4 is worth naming, because it is the failure mode this whole set is supposed to guard against: **the prose acknowledged a weaker guarantee while the headline invariant continued to assert the stronger one.** That is assurance-label imprecision, and it is exactly what Annex B exists to prevent.

What changed in v0.5.3. A third external validation pass produced three corrections.

1. **Article 99(6a) is restored, and correctly.** Two earlier passes described the small mid-cap cap differently and I twice declined to print either reading. The reviewer supplied the Official Journal reference, and a second independent source confirms it: the cap reaches **Art. 99(4) and 99(5) only**, not the Art. 99(3) prohibition band. The practical consequence is now stated — a small mid-cap that breaches an Article 5 prohibition faces the full 35 million euro or 7 % band, unlike an SME. The v0.5.2 note calling the mechanism unverifiable is withdrawn; it was accurate about my own sourcing at the time and wrong about the law.
2. **The one-page and executive-summary wording no longer asserts preventive stale rejection unconditionally.** §4.4 had already split GI-6 into GI-6P and GI-6D, but the front of the paper still said an approval that goes stale "must be re-evaluated or re-approved" and told a board to measure "rejection of stale approvals" — both true only under GI-6P. The strong guarantee was being reintroduced two pages before it was withdrawn.
3. **"Not assessed" is a status, not a class** (§2.5). The companion specification briefly carried a sixth provenance class, PX, which left this set with two taxonomies. It is withdrawn there and here: the canonical taxonomy is P0 to P4 in all three documents, and an unassessed parameter carries no class at all.

What changed in v0.5.4. Two corrections, both small, and one of them is a matter of the paper's own evidence policy rather than of law.

1. **§3.6 said that the absence of a conditional-write primitive is "the common case across major enterprise platforms".** The paper's Evidence Status states that market-wide claims are hypotheses unless supported by published evidence, and no adapter survey has been published. The claim is now that primitive availability is a property of the specific API and resource, to be tested rather than assumed — which is also what §3.6 already told implementers to do two sentences later.
2. **Companion references** point at GAIP 0.4 and Guide 0.1.4.

Open corrections carried forward. The Estonian statutory citations that supported the stewardship chapter were verified against two independent mirrors rather than against the live consolidated text at riigiteataja.ee, which serves its English texts through JavaScript. Those section numbers should be checked against the live consolidation before the stewardship document

is published.

Imprint and contact

Publisher and licensor

GoSec Cloud OÜ 9 Järvevana tee, 11314 Tallinn, Estonia Estonian Commercial Register no. 17359627 · VAT EE102990704 Registry court: Tartu District Court Authorised representative: Yves Frédéric N'Soussoula info@gosec.cloud · +372 60 26553

Vianordis is a product of GoSec Cloud OÜ and is not itself a legal entity.

Author

Yves Frédéric N'Soussoula, Founder & CEO, GoSec Cloud OÜ.

Feedback on this paper — Open Review

Substantive criticism, corrections and counter-arguments are welcome through https://www.vianordis.eu/en/contact. Please indicate the section number. Corrections of fact will be applied in the next edition and listed in Annex D; substantive critique will be credited unless you prefer otherwise.

Related routes: platform and applications https://www.vianordis.eu/en/services · release status https://www.vianordis.eu/en/release-status · source and licensing https://www.vianordis.eu/en/git · trust centre https://www.vianordis.eu/en/trust · community and contribution https://www.vianordis.eu/en/build-the-constellation

Companion documents

Vianordis GAIP Candidate Specification 0.4 — the normative technical profile. *Vianordis Adoption & Measurement Guide 0.1.4* — worked example, adoption path, maturity model and measurement protocol.

Document licence

Cite as

N'Soussoula, Y. F. (2026). *The AI Control Gap: Why approving a model is not approving an action*. Vianordis Position Paper 01, version 0.5.4. Tallinn: GoSec Cloud OÜ. https://www.vianordis.eu

Legal status: law as at 26 August 2026. Not legal advice.

End of draft — version 0.5.4