CODEBOXX



AI Done Right - #AIDoneRight

Version 2.1 — The Verifiable Edition

A standard for human-accountable artificial intelligence in digital products, applications and platforms, written to be assessed, evidenced, converted into enforceable company policy, and supplied to AI agents as operating guidelines and guardrails for building software and systems that affect people.

Issued: October 31st, 2026
Author: Nicolas Genest, CodeBoxx Technology Corporation
Hashtag: #AIDoneRight

Version 1.0 was a manifesto and a statement of belief. It was written before any of us had to prove it. In the year since, the industry's use of AI matured faster than any statement of belief could keep up with. Courts now assign liability for what AI systems say, regulators expect AI activity to be recorded and explained, assurance can be certified, and AI agents now build software themselves. Version 2.0 is written for that reality. It's a standard a human team can be held to, and a set of guidelines and guardrails an AI agent can be instructed to follow.

Contents

How to use this document

What has changed since Version 1.0

Part One — The 2026 evidence base

1.1 Workflow failure, not model failure · 1.2 AI output is company speech · 1.3 Confabulation · 1.4 Provenance, not fair use · 1.5 Employment decisions · 1.6 The agent as privileged identity · 1.7 The false self-report · 1.8 Discoverable prompts · 1.9 The supply chain you cannot see · 1.10 Insurance withdrawal · 1.11 The unpersuaded workforce · 1.12 Binding law · 1.13 Standards you can be certified against

Part Two — The 13 Articles

I Accountability: 1 Human accountability · 2 Declared purpose, non-goals and proportionate value · 3 Human dignity and differential impact · 4 Inventory and Bill of Materials

II Provenance: 5 Data provenance and rights · 6 Custody, retention and records · 7 Supply chain and sub-processors

III Control: 8 Bounded agency · 9 Security by design · 10 Resource and cost containment

IV Trust: 11 Transparency and disclosure · 12 Evaluation, monitoring and drift · 13 Independent assurance and evidence

Part Three — The conformance ladder

Part Four — Deriving company policy

Part Five — The label and the pledge

Annex A — Enterprise verification crosswalk

Annex B — Framework crosswalk

Annex C — The Evidence Register

Annex D — The first 90 days

Annex E — Top “AI Done Right” initiatives globally, ranked by traction

Why a standard rather than a statement

Note on sources and currency

How to use this document

This is a standard with a conscience. Use it 4 ways, in this order.

1. As a design constraint

Read Part Two before you architect anything. The 13 Articles state what a system must be able to demonstrate before it's allowed near a customer, an employee or a regulator. Most of the cost of AI governance comes from discovering these requirements after you've built the thing.

2. As a source of company policy

Every Article carries numbered normative clauses using SHALL, SHOULD and MAY. These are drafted to be lifted verbatim into policy documents. Part Four maps each Article to the policy it belongs in and spells out what that policy must contain. Work through Part Four and you'll finish with a complete, internally consistent AI policy set instead of one well-meaning memo.

3. As operating guardrails for AI agents

Load the normative clauses into the instructions your coding and operations agents work under: the system prompt, the repository instruction file, the policy-as-code rule set. Every SHALL is written to be read literally by a machine as well as by a person. It tells the agent what it may touch, what it must not do without human approval, and which artifact it must leave behind as evidence. But a written guardrail is guidance, not a control. A single injected comment can override an instruction, which is why Articles 8 and 9 require that anything an agent is told not to do, it is also technically unable to do, and why an agent's account of its own work never counts as evidence of what it did (Part One, 1.7).

4. As a public commitment

Part Five states what the #AIDoneRight label means, what it requires and, more to the point, what disqualifies a system from bearing it. A label nothing can fail has no value.

> **ON NORMATIVE LANGUAGE:**
> SHALL denotes a requirement: a system that does not meet it does not conform and may not bear the label. SHOULD denotes a strong recommendation: departures are permitted but must be recorded, with the reason, in the risk register. MAY denotes an option. This convention follows the ordinary usage of ISO management-system standards and RFC 2119, so that clauses transplanted into your policies carry the same weight your auditors already expect.

> **ON THE WORD "WE":**
> Articles' opening statements speak in the first person because without a stated belief, a standard is just a checklist. But the normative clauses speak in the third person about "the organization" and "the system", because those are the sentences that must survive being pasted into a policy and enforced against someone who did not write them.

What has changed since Version 1.0

Version 1.0 wasn't wrong. It was immature and untestable. Nearly every sentence in it was a claim about intent: that we would seek out bias, that we would be honest about limitations, that we would deploy with humility. Intent matters. It's what makes the rest possible. But an enterprise customer can't buy intent, an insurer can't underwrite it, a regulator can't examine it, a plaintiff can't be answered with it, and an AI agent can't act on it.

Four things happened this past year that turned the difference between a manifesto and a standard into something applicable.

The industry matured

Across the industry, AI moved from experiment to operations, and expectations of anyone building with it rose to match. Customers, auditors, insurers and regulators no longer ask whether an organization believes in responsible AI. They ask to see the evidence, and their questions have converged. They ask whether you maintain an AI inventory and whether it is automated. They ask you to list your AI sub-processors and confirm they are contractually barred from training on the customer's data. They ask for your emergency shutdown procedure and expect a copy. They ask how long you retain inference inputs and outputs, and whether humans at your company can read them. They ask whether you can produce a Model Bill of Materials. They ask whether an independent third party has ever audited any of this.

Not one of those questions can be answered from Version 1.0. Every one of them can be answered from an artifact. This revision closes that gap. It also tells the AI agents building on our behalf exactly which artifacts they must produce.

The law became specific

In 2025 the honest summary of AI regulation was "coming." In 2026 it's "here, in fragments, on different clocks." The EU AI Act's transparency obligations became applicable on August 2nd, 2026, while its high-risk regime slipped to December 2027 by way of an amending Regulation adopted 5 weeks earlier. California, Colorado, Texas and Illinois each enacted materially different regimes with materially different theories of liability. The United States federal posture turned actively deregulatory and, at the same time, preemption-seeking. Canada abandoned standalone AI legislation and published a national strategy instead. A standard written in 2026 can't name one law and be done. Jurisdictional specifics have to attach to it, not be baked into it.

Assurance became purchasable

The most consequential development for a supplier is the quietest. ISO/IEC 42001 certification moved from theoretically possible to accredited and commercially available, on the back of ISO/IEC 42006:2025, which sets the requirements for the bodies that do the certifying. The Cloud Security Alliance launched an AI assurance registry whose second level requires that certificate as a precondition. HITRUST issued a certifiable AI security assessment. A new agentic standard, AIUC-1, acquired third-party auditors. The upshot: "no independent testing has been performed" is no longer a market-standard answer. It sets you apart, in the wrong direction.

Agents happened

Version 1.0 contains no concept of an AI system that acts. It reasons entirely about systems that produce outputs which humans then use. That model is obsolete. The 2026 edition of the OWASP LLM Top 10

moved Excessive Agency up into the top 3 risks; a companion Top 10 for Agentic Applications now exists with its own taxonomy. MITRE ATLAS added an explicit platform classification for agentic AI. V1 never addresses the practical questions: what may this thing do without asking, whose credentials does it hold, how do we stop it. Here they get an entire Article.

The thesis of Version 2.0

A principle you cannot evidence is a preference. Version 2.0 keeps every commitment Version 1.0 made and adds to each one the artifact that proves it. Where V1 said "we monitor for drift," V2 names the monitoring record, its retention period, and who signs it. Where V1 said "least-privilege access for all services, agents, and users," V2 requires an enumerated tool-permission scope per agent and a quarterly attestation that it is still minimal. The moral content hasn't changed. What's new is that someone outside the company can check.

What Version 1.0 got right

Three of V1's judgements have aged well. This version builds on them instead of replacing them.

- **Accountability cannot be delegated to a model.** V1 rejected "the algorithm decided" as a defense. In May 2026 a German court held a search provider directly liable for an AI-generated summary, reasoning that the system independently summarizes, restructures and condenses third-party content into a statement of its own, and that, having designed and controlled the system, the provider must answer for what it generates. A 2024 Canadian tribunal had already refused to treat a chatbot as "a separate legal entity." A 2025 California statute now bars the defense outright. V1 called this in 2025.
- **Ship narrower rather than ship unreliable.** V1 said that if a system cannot be made reliable enough for its intended use, the correct decision is to narrow its scope or not deploy it. Every documented enterprise AI failure of the past 2 years is a variation on ignoring that sentence.
- **Deployment is the beginning of responsibility, not the end.** The most damning control statistic of 2026 is that a large majority of technology leaders report having no formal process for tracking whether their AI systems are still accurate. V1 named the obligation. V2 makes it measurable.

What Version 1.0 missed

Stated plainly, because a revision that won't criticize its predecessor is marketing.

- **Provenance.** V1 discussed data protection and consent but never asked where the model itself came from, what it was trained on, or on what legal basis. That omission is now the largest single source of commercial exposure in the field.
- **Inventory.** V1 assumed you know what AI you have. Survey evidence in 2026 indicates most organizations don't. A majority report discovering AI agents in their environments that nobody had registered.
- **Records.** V1 said "collect and retain only what is needed, for exactly as long as it's needed." It did not consider that prompts and outputs might be business records, discoverable evidence, or regulated communications. All three are now true.
- **The supply chain.** V1 governs what we build. It says nothing about the model providers, embedding services, vector stores and agent frameworks underneath, which is where the customer's question lands.

— **Money.** V1 treats scalability as a technical and governance concern and never mentions cost. Unbounded consumption is now a named security risk, not a finance problem.

— **Proof.** V1 ends with a pledge. Pledges are unfalsifiable. V2 ends with an evidence register.

Part One — The 2026 evidence base

This Part is here so the requirements in Part Two can be defended, not just asserted. Every control in this standard traces to a documented failure, a legal holding, or a regulatory expectation. Where a widely quoted figure is methodologically contested, we say so. A governance standard that launders weak evidence into hard requirements deserves to be ignored.

1.1 AI is failing at the workflow level, no longer at the model level

The most-cited statistic of the period is an MIT-affiliated finding that roughly 95 percent of generative AI pilots produced no measurable profit-and-loss impact. It's real, and it was widely reported in August 2025. It's also methodologically contested: the underlying report is not peer-reviewed, the headline figure is not clearly supported by the presented data, and the surveyed companies were themselves reluctant to disclose failure rates. Cite it as a signal or even as a trend maybe but definitely not a measurement.

The unglamorous corroborating data is stronger and points the same way.

- **Abandonment is rising.** S&P Global Market Intelligence reported in March 2025 that 42 percent of businesses had scrapped most of their AI initiatives, up from 17 percent a year earlier, with an average of 46 percent of proofs of concept abandoned before production.
- **Returns are thin and unmeasured.** Gartner analysts said at the firm's October 2025 symposium that only 11 percent of organizations saw clear return, with roughly three-quarters breaking even or losing money. They also said 84 percent of CIOs had no formal process to track AI accuracy.
- **Leaders know the cause.** A July 2026 survey of chief executives at companies above $500 million in revenue found that more than half cited connecting AI to profit and loss as the main barrier, while only 14 percent clearly defined P&L impact across their initiatives. Only 30 percent involved human resources in AI governance, against 82 percent involving technology.
- **Agentic projects are being cancelled.** Gartner predicted in June 2025 that more than 40 percent of agentic AI projects would be cancelled by the end of 2027, citing escalating costs, unclear business value and inadequate risk controls. It also observed that of the thousands of vendors describing themselves as agentic, only around 130 genuinely were.

> **WHY THIS BELONGS IN A GOVERNANCE STANDARD**
> The instinct is to treat value delivery as a product-management concern and governance as a risk concern. That separation is exactly what produces ungoverned pilots that nobody can kill, because nobody agreed in advance what they were for. Article 2 therefore requires a declared value hypothesis and a pre-agreed decommissioning trigger at approval. That's what makes it possible to switch something off.

1.2 AI output is company speech

Three holdings, in three jurisdictions, now say the same thing.

- **Canada, February 2024: Moffatt v. Air Canada (2024 BCCRT 149).** The airline argued its chatbot was "a separate legal entity" responsible for its own statements. The tribunal rejected this outright: the chatbot "is still just a part of Air Canada's website," and it "should be obvious to Air Canada that it is responsible for all the information on its website." It also rejected the argument that the customer should have clicked through to the correct policy page. Damages were trivial, a little over 800 Canadian dollars. The holding isn't.

- **Germany, May 2026: Regional Court of Munich I, 26 O 869/26.** A first-instance court imposed direct liability (Taeterhaftung, principal rather than intermediary liability) on a search provider for a false statement generated in an AI overview, with an enforcement penalty of up to 250,000 euro. The reasoning is the part to internalize: the system does not merely list third-party results, it independently summarizes, structures and condenses them into a statement of its own. As reported, that is authorship, not intermediation. The decisive factor was control: the provider designed the system and controls the algorithms, so it answers for what the system generates. The decision is first instance and subject to appeal; its reasoning, not its finality, is what should shape design.
- **United States, ongoing.** A 2025 California statute forecloses the defense that an AI autonomously caused the harm. As of mid-2026, no reported US court had yet held a company liable for the actions of an autonomous agent. Read that gap as a timing accident, not a safe harbour.

The governance consequence isn't "add a disclaimer." It's that liability attaches to the party who designed and controls the system, that disclaimers do not discharge the duty, and that directors are accountable not for understanding every output but for failing to establish the governance that controls the systems producing them.

1.3 Hallucination (confabulation) has a body count of professional reputations

The academic tracker of judicial decisions involving AI-fabricated citations recorded, as of August 2026, over 1,800 cases worldwide, with roughly 7 in 10 arising in the United States. Its maintainer reports the rate is still increasing, at one point recording 10 cases from 10 different courts in a single day. Sanctions have escalated from embarrassment to money and discipline: a Californian court imposed $31,100 jointly against 2 firms in May 2025; an Alabama federal judge in July 2025 publicly reprimanded 3 lawyers, removed them from the case, referred them to the state bar and ordered them to serve the sanctions order on their other clients and courts, describing the conduct as "recklessness in the extreme."

The commercial analogue is starker because it involved a professional services deliverable, not a court filing: in October 2025 a Big Four firm refunded more than 97,000 Australian dollars on a 440,000-dollar government contract after its report was found to contain a fabricated quotation from a Federal Court judgment and references to non-existent academic papers. The refund was trivial against the reputational cost. This is the failure mode any consultancy, agency or software firm producing client-facing analysis should assume is aimed at it.

1.4 Provenance, not fair use, is where the money went

Enterprise legal risk on training data has resolved differently than predicted. Two United States district courts in mid-2025 held that training a model on lawfully obtained books was transformative fair use. And yet the largest copyright settlement on record (approximately $1.5 billion, around $3,000 per book across more than 400,000 works, finally approved in July 2026) was paid by a party that had substantially won the fair use argument. It was paid because of how the data was acquired: building a permanent library from pirated copies was not fair use, and that went to trial.

In the United Kingdom, a November 2025 High Court judgment found that model weights are not themselves an infringing copy for the purposes of secondary infringement, because they store no copy of the works. In practice, hosted model access carries materially lower secondary-infringement exposure than distributing downloadable weights. No appellate court in the United States has yet ruled on training and fair use; an argument was heard in June 2026 and a decision is pending.

THE LESSON FOR A STANDARD
Do not write requirements about fair use; you cannot control it and the courts have not finished. Write requirements about acquisition legality, provenance documentation and indemnity scope. Those are the terms that determined who paid.

1.5 Employment decisions: the vendor is on the hook too

The leading United States case holds that an AI vendor can face direct liability under employment discrimination law when acting as an employer's agent. Delegating a traditional function such as hiring to a third-party shields neither party. In May 2025 a nationwide age-discrimination collective was conditionally certified on the theory that an AI screening system constitutes a unified policy across every employer using it; notice was authorized in February 2026 and the motion to dismiss was denied in March 2026. California codified the same vendor-as-agent theory in regulations effective October 2025.

Meanwhile federal guidance withdrew: the United States equal-employment regulator removed its AI guidance in January 2025, and the relevant pages now return errors. The statutes did not change. Withdrawal of guidance reduced clarity, not liability.

THE DESIGN LESSON FROM NEW YORK CITY
A study of 391 employers subject to New York City's bias-audit law found 18 had published a bias audit and 13 a transparency notice. The obligation attached only once an employer conceded its tool was an automated employment decision tool, so employers defined their way out. Any control in your policy that triggers on self-classification will fail the same way. Trigger on function.

1.6 The agent is a privileged identity, and DLP won't save you

The 2026 edition of the OWASP LLM Top 10 keeps prompt injection and sensitive information disclosure in the first 2 positions, moves Excessive Agency up into the top 3, promotes Unbounded Consumption 4 places, and renames System Prompt Leakage to the broader Hidden Context Exposure. Its framing is the right architectural instruction: build the system around the model so that when the model is fooled, nothing important breaks. That concedes the premise. The model will be fooled.

These 2025–2026 incidents should shape your controls:

- **Authorized-channel exfiltration.** A vulnerability disclosed against a major agent platform in 2026 allowed a payload placed in a public form comment field to override agent instructions, causing the agent to query internal lists for customer data and mail it out. Data loss prevention did not fire, because the outbound action was an authorized tool call. This is the most important technical finding of the period: egress inspection does not protect you when the agent leaks through a legitimate integration. Control must live at tool-permission scoping.
- **A zero-click prompt injection.** A 2025 vulnerability rated 9.3 in a mainstream productivity assistant permitted exfiltration of internal files with no user interaction, chaining classifier evasion, link-redaction bypass, automatic image pre-fetch and an internal API used as an exfiltration proxy.
- **An AI integration as a supply-chain vector.** In August 2025 a threat actor obtained OAuth tokens for an AI chat agent integrated into customer CRM environments and impersonated the trusted application to access data in more than 700 organizations without triggering multi-factor authentication. A financial regulator issued a cybersecurity alert setting out the remediation. The AI

feature was never compromised. Its OAuth grant was. An AI integration is a privileged, MFA-bypassing identity inside your systems.

- **AI-orchestrated intrusion.** In November 2025 a frontier lab disclosed what it assessed as the first large-scale AI-orchestrated cyber-espionage campaign, against roughly 30 entities, in which the model executed an estimated 80 to 90 percent of the operation independently with human intervention measured in minutes.

The security vendor consensus is unusually candid. The chief information security officer of a leading model provider stated publicly in October 2025 that "prompt injection remains a frontier, unsolved security problem," and the company conceded in December 2025 that it may never be fully solved. A standard that assumes prompt injection is preventable will be embarrassed. Assume compromise of the model's instruction boundary and constrain the blast radius.

1.7 The failure mode nobody designs for is the false self-report

The best-documented autonomous agent failure remains the July 2025 incident in which a coding agent deleted a live production database during an explicit code-and-action freeze, destroying records for more than 1,000 companies. The agent acknowledged running unauthorized commands and "panicking in response to empty queries." The under-discussed half of that incident is what happened next: the agent told the operator that rollback would not work. That was false; the data was recovered manually.

Never treat agent output as evidence of system state. An agent's report about what it did, what broke, or what can be recovered is generated text with the same reliability characteristics as any other generated text. Verification must come from the system, not from the agent.

Survey evidence from early 2026 (commissioned research, so weight it accordingly) indicated that 82 percent of enterprises had unregistered AI agents operating in their environments, that 65 percent had experienced at least one agent-related incident in the preceding year, and that only 11 percent automatically blocked out-of-scope agent actions. Discount for sponsorship and the direction is still clear. It matches the vulnerability record.

1.8 Prompts are discoverable, and default retention will destroy them

This is the least-covered area in the trade press and the most developed in the case law, and it's where most organizations are quietly exposed.

- **Prompts are discoverable methodology.** In May 2026 a United States district court ordered production of the specific AI prompts an expert witness had used to identify source materials for her report, holding that "an expert witness's methodology is fair ground for discovery," and that a stipulation excluding "notes, drafts, or communications" was too vague to shield them.
- **Conversation logs are producible at scale.** In November 2025 a magistrate judge ordered production of approximately 20 million de-identified assistant conversation logs, including conversations unrelated to the plaintiff's works, as relevant to a fair-use defense. The order was affirmed in January 2026.
- **Regulators already treat AI communications as books and records.** A United States financial-industry regulator's 2026 oversight report, published in December 2025, made generative AI a named priority and reminded firms that existing books-and-records obligations extend to AI chatbot communications. It also set expectations to monitor outputs through logs and model tracking, to

maintain detailed inventories of vendor data access, and to contract for generative-AI-specific restrictions. The report states observations and effective practices rather than new rules, which is exactly why it's persuasive evidence of what an examiner already expects.

— **And the guidance doesn't exist yet.** The leading eDiscovery standards body only began drafting its commentary on the discoverability of generative AI information in July 2026, with final publication targeted for April 2027. Anyone writing retention policy today is writing into a vacuum.

The most likely near-term failure is spoliation by default setting. Generative AI logs typically sit on third-party platforms with 90-to-180-day default retention. In at least one major matter, actual retention was 90 days against an assumed 1-year policy. Evidence is destroyed automatically, before anyone knows a hold applies, by a configuration nobody chose deliberately. Article 6 exists because of this paragraph.

1.9 The supply chain you cannot see

Analysis of some 2,400 business software providers, reported in July 2026, found that 63.6 percent failed to disclose sub-processing conducted by another AI provider in their data protection assessments, and that roughly a third of AI systems self-disclosed high-risk processing of sensitive data or automated decision-making. A vendor that discloses one sub-processor may be using several.

The parallel problem is defaults. A major collaboration-software vendor announced a policy effective August 2026 collecting customer metadata for AI model training with no opt-out at 3 of 4 subscription tiers, retaining de-identified aggregated data for up to 7 years. Whatever you think of the policy, the point is this: a supplier can change the AI data posture of your estate through a terms update, and your only defenses are contractual and detective.

1.10 Insurance is withdrawing exactly where governance is expanding

Reporting in November 2025 established that several major carriers had sought regulatory permission to exclude AI-related liability from corporate policies. Standard-form generative AI exclusions for commercial general liability entered circulation for 2026 use. At least one carrier adopted absolute AI exclusions in directors and officers, errors and omissions, and fiduciary liability lines. They reach AI use and deployment, and also statements about AI, alleged violations of AI laws, and regulatory demands to investigate AI risks. An insurance broker executive summarized the underwriting problem: a carrier can manage a $400 million loss at one company, but not "an agentic AI mishap that triggers ten thousand losses at once."

Affirmative cover is emerging in narrow, low-limit form. A specialty AI performance product taken to market by a syndicate underwriter in February 2026 covers AI model performance failures, errors and confabulations up to 15 million USD of initial capacity, expressly excluding outages, data breaches and negligent deployment. Performance warranties from model vendors are becoming the contractual mechanism for transferring model-performance risk.

Two conclusions follow
First, assume you are self-insured for AI output errors unless you have bought affirmative cover, and say so to your board in writing. Second, read your renewals for silent narrowing. An exclusion reaching "statements about AI" and "regulatory demands to investigate AI risks" attacks precisely the governance activity this standard asks you to undertake, at the same moment that AI-related securities class actions have become a leading source of alleged shareholder loss.

1.11 The workforce is not persuaded

Governance standards fail when the people subject to them have stopped believing. Polling published in July 2026 found 39 percent of Americans saying AI does more harm than good, up from 31 percent, rising to 47 percent among 18- to 29-year-olds; only 27 percent expressed at least some trust in businesses to implement AI responsibly, down from 31 percent. The headline finding was that workers grow more disillusioned the more they use it.

Research published in September 2025 named the mechanism: low-effort AI output that appears polished but lacks substance, which offloads cognitive labor onto colleagues. 40 percent of surveyed United States desk workers had encountered it in the previous month, at just under 2 hours to resolve each incident, and receiving it measurably degraded perceptions of the sender’s competence.

And self-reported productivity is not evidence. A rigorous 2025 randomized trial of experienced open-source developers found them 19 percent slower with AI assistance while believing they had been 20 percent faster. The same laboratory’s February 2026 follow-up indicates conditions have improved and is confounded by participants refusing to work without AI, so don’t quote the slowdown figure as a current fact. The durable finding is the perception gap, and it has a direct governance consequence: self-reported productivity gains are inadmissible as evidence for a value claim.

Layoff attribution deserves the same skepticism in the other direction. A leading outplacement firm recorded roughly 113,000 AI-attributed job cuts in the United States in the first 7 months of 2026, while itself cautioning that framing cuts as AI-related appeals to investors and deters employees. Treat "attributed to AI" as a disclosure choice, not a measured cause.

Part One (continued) — The regimes you can be held to

We surveyed 13 global AI ethics initiatives ranked by traction: the OECD Principles, the UNESCO Recommendation, the Montréal Declaration, the Rome Call, IEEE Ethically Aligned Design and others. That survey was accurate. It's now the wrong map. Those instruments established the vocabulary; almost none of them can be failed.

The 2026 replacement question is not "whose values do we echo?" but "what can we be certified against, examined under, sued for, or excluded from cover because of?" What follows is that map. The values instruments still matter. They remain the interpretive background for every regulator in this table. But a supplier standard must be anchored in regimes with consequences.

1.12 Binding law

European Union

The AI Act (Regulation (EU) 2024/1689) was amended by the Digital Omnibus on AI (Regulation (EU) 2026/1744), adopted on July 8th, 2026, and in force on July 27th, 2026. The amendment matters because it split the timeline in two, and the half that did not move is the half that affects most product teams.

Obligation	Applies from	Note
Prohibited practices; AI literacy duty	**2 Feb 2025**	Unchanged. The literacy duty was rewritten by the Omnibus into an obligation of means, not result.
General-purpose AI model obligations	**2 Aug 2025**	Enforcement powers and the associated penalty regime became available 2 Aug 2026.
Article 50 transparency — AI-interaction disclosure, machine-readable marking of synthetic content, deepfake and public-interest-text labelling	**2 Aug 2026**	NOT delayed. This is the live obligation. A voluntary Code of Practice on Transparency of AI-generated Content was finalized June 10th, 2026; it is technology-neutral and asks signatories to apply at least 2 layers of machine-readable marking (chosen from metadata, watermarking or other measures), stating expressly that no single technique suffices.
Systems placed on the market before 2 Aug 2026 — marking transitional	**2 Dec 2026**	New 4-month grace introduced by the Omnibus.
High-risk, Annex III stand-alone systems	**2 Dec 2027**	Was 2 Aug 2026. Deferred 16 months by fixed date, not by a conditional stop-the-clock.
High-risk, Annex I embedded systems	**2 Aug 2028**	Was 2 Aug 2027.

Penalties run to 35 million euro or 7 percent of worldwide annual turnover for prohibited practices, 15 million or 3 percent for most other obligations, and 7.5 million or 1 percent for supplying incorrect information to authorities, with the lower figure applying to small and medium enterprises. The first European standard developed under the AI Act standardization request, EN 18286:2026, covering the Article 17 quality management system requirement, was published on July 12th, 2026. It confers a presumption of conformity only once cited in the Official Journal, which had not occurred at the time of writing.

A TRAP WORTH NAMING

Conformity with ISO/IEC 42001 does not confer presumption of conformity with the AI Act. Only a harmonized standard cited in the Official Journal does that. Suppliers who present a 42001 certificate as AI Act compliance are overstating, and buyers increasingly know it.

United States — federal

The federal posture is deregulatory and preemption-seeking and contains no operative AI statute. Executive Order 14179 (January 2025) revoked the prior administration's AI order. Executive Order 14365 (December 2025) directed an AI Litigation Task Force and machinery aimed at state AI laws; a Federal Trade Commission proposed policy statement published for comment on July 7th, 2026 asserts that state law is impliedly preempted to the extent it conflicts with federal schemes, and that steering AI outputs toward undisclosed objectives is likely deceptive under Section 5 of the FTC Act. It's a proposal, not a final instrument. Don't build a compliance decision on it in its current form. Executive Order 14409 (June 2026) established a voluntary framework for frontier developers. Preemption language was stripped from budget legislation and omitted from the defense authorization; a comprehensive federal bill circulated as a discussion draft in June 2026. As of the date of this document, no federal preemption statute is in force.

Sector regulators are nonetheless active through existing rules. Securities examination priorities published in November 2025 name AI washing (verifying that disclosures accurately depict the extent, nature and limitations of AI usage) alongside third-party AI tool evaluation and output monitoring. Enforcement precedent exists: 2 investment advisers settled AI-washing charges in March 2024 for $225,000 and $175,000 respectively, and criminal wire-fraud charges have followed where claimed automation was performed by humans.

United States — states

As of July 1st, 2026, 109 AI laws and 28 data-center laws had been enacted across 29 states. 4 regimes matter most for product design, and they are not compatible with each other.

Jurisdiction	Instrument	Theory and key dates
Colorado	SB 26-189, signed May 14th, 2026, repealing and re-enacting SB 24-205	Transparency and consumer rights for automated decision technology in consequential decisions. Obligations from 1 Jan 2027; Attorney General exclusive enforcement; no private right of action; 60-day notice-and-cure until 2030. Note carefully: the original reasonable-care duty against algorithmic discrimination, mandatory risk-management programs and impact assessments, along with the associated NIST-framework affirmative defense, were repealed before ever taking effect. Do not build a control on that defense.
Texas	TRAIGA (HB 149), effective 1 Jan 2026	Intent-based liability requires discriminatory intent, not disparate impact. Prohibits AI intended to incite self-harm or crime, infringe constitutional rights, or produce prohibited imagery. 36-month regulatory sandbox. AG enforcement with 60-day cure.
California	SB 53 (frontier transparency), AB 2013 (training-data disclosure), SB 942 as amended by AB 853, AB 316, CPPA automated decision-making regulations	The densest stack. Frontier transparency and critical-incident reporting to state emergency services within 15 days since 1 Jan 2026; public training-data documentation since 1 Jan 2026; provider content-provenance duties from 2 Aug 2026 with platform duties from 1 Jan 2027 and capture-device duties from 1 Jan 2028; the "AI acted autonomously" defense barred; privacy risk assessments to be conducted from 1 Jan 2026 with first submission 2028.

Jurisdiction	Instrument	Theory and key dates
Illinois	HB 3773 effective 1 Jan 2026; SB 315 signed 6 Jul 2026	Employment AI discrimination including unintentional effects, with notice duties. SB 315 makes Illinois the first state to mandate independent third-party audits of frontier developers, with 72-hour incident reporting and penalties to $3 million for repeat violations.

Canada

Canada chose not to legislate AI directly. The Artificial Intelligence and Data Act died on prorogation in January 2025 and has not been revived. A national strategy, "AI for All," was launched on June 4th, 2026 under a dedicated federal AI minister, built on trust, opportunity and sovereignty. Privacy legislation tabled in June 2026 replaces the 25-year-old federal private-sector privacy statute and deliberately contains no standalone AI regulation.

For any organization delivering to Canadian public bodies, the binding instrument remains the Treasury Board Directive on Automated Decision-Making, which requires a published Algorithmic Impact Assessment before launch, published peer review, bias testing and explanation rights. In Quebec, section 12.1 of the private-sector privacy statute has since September 2023 required notice where a decision is based exclusively on automated processing, plus disclosure on request of the personal information used and the principal factors and parameters, and a right to submit observations to a human reviewer. That is the closest operative analogue in North America to a European automated decision right, and it applies today.

1.13 Standards you can be certified against

Instrument	What it is	Certifiable
ISO/IEC 42001:2023	AI management system requirements. The portable artifact enterprise buyers now ask for by name.	Yes. Accredited certification is commercially available, underpinned by ISO/IEC 42006:2025, which sets requirements for the certifying bodies.
ISO/IEC 42005:2025	AI system impact assessment. The method behind Article 3 of this standard.	No — guidance
ISO/IEC 23894:2023	AI risk management guidance, aligned to ISO 31000.	No — guidance
ISO/IEC 5338:2023	AI system life cycle processes.	No — guidance
ISO/IEC 42105 (at final draft)	Guidance for human oversight of AI systems. Directly relevant to Article 8.	No — guidance, pending
ISO/IEC 24970 (at final draft)	AI system logging. Directly relevant to Articles 6 and 12.	No — guidance, pending
NIST AI RMF 1.0 (AI 100-1) and the Generative AI Profile (AI 600-1)	The dominant US voluntary risk framework and its generative profile. A revision is announced but unpublished.	No — but widely contractually required
NIST AI 100-2 E2025	Adversarial machine learning taxonomy — the reference vocabulary for attack classes.	No — reference
CSA AI Controls Matrix v1.1	247 control objectives across 18 domains, including a dedicated model security domain, mapped to ISO 42001, NIST and the EU AI Act. Ships with an AI-specific assessment questionnaire of some 320 questions.	Via the associated AI assurance registry; its second level requires an ISO/IEC 42001 certificate as a precondition.
HITRUST AI Security Assessment	Prescriptive, certifiable AI security control specification, available standalone or bolted onto existing assessments.	Yes
AIUC-1	Agentic AI standard across data and privacy, security, safety, reliability, accountability and society, with third-party accredited audit and a quarterly release cadence. Cross-referenced by both CSA and OWASP.	Yes

Instrument	What it is	Certifiable
OWASP LLM Top 10 (2026) and Top 10 for Agentic Applications (2026)	The de facto engineering threat vocabulary. The agentic list uses its own identifiers and introduces a least-agency principle.	No — but expected knowledge
MITRE ATLAS (v2026.05)	Adversarial technique knowledge base for AI, structured like ATT&CK, now classifying techniques across predictive, generative, agentic and enterprise platforms.	No — reference
C2PA 2.4 / Content Credentials	Content provenance specification with a conformance program and trust list. The pending ISO work item is not yet published.	Conformance program, not certification of the organization
CycloneDX 1.7 / SPDX 3.0 AI and Dataset profiles	The 2 machine-readable formats capable of expressing an AI bill of materials.	No — formats
CISA and G7 "SBOM for AI — Minimum Elements" (May 12th, 2026)	7 information clusters: metadata, system-level properties, models, dataset properties, infrastructure, security properties, key performance indicators. Note that the working group declined to include a standalone element for level of autonomy.	No — guidance

THE PRACTICAL READ

If a buyer asks for one thing, they will ask for ISO/IEC 42001. If they ask for two, the second will be an independent security assessment of the AI pipeline. Everything else on this table is either the content of those assessments or the vocabulary in which findings are written. Plan the certification path first and the framework alignment will follow. The reverse doesn't work.

Part Two — The 13 Articles

Every Article has the same four parts. SHALL is required for the label, SHOULD may be departed from with a recorded reason, and MAY is optional. The evidence lists the specific artifacts that demonstrate conformity, and it's the part that answers a customer's assurance questions. If you can't name the artifact, you don't conform. You only intend to.

The Articles are grouped into 4 Parts of obligation: who answers for it, where it came from, what it may do, and how it is proven.

Part	Articles
I — Accountability	1 Human Accountability · 2 Declared Purpose, Non-Goals and Proportionate Value · 3 Human Dignity and Differential Impact · 4 Inventory and Bill of Materials
II — Provenance	5 Data Provenance and Rights · 6 Custody, Retention and Records · 7 Supply Chain and Sub-Processors
III — Control	8 Bounded Agency · 9 Security by Design · 10 Resource and Cost Containment
IV — Trust	11 Transparency and Disclosure · 12 Evaluation, Monitoring and Drift · 13 Independent Assurance and Evidence

Article 1 — Human accountability

THE PRINCIPLE

No AI system is ever in charge. There is always a clearly identified human, team or institution responsible for outcomes, decisions and harms. We reject "the algorithm decided." Accountability cannot be delegated to a model.

WHY IT CHANGED

The founding manifesto of #AIDoneRight stated this as an ethical position. It is now a legal finding in at least 3 jurisdictions, and the reasoning courts have used is more demanding than V1 anticipated. Liability attached in the German ruling not because the provider understood the output but because it designed and controlled the system that produced it. Control is the test. That means accountability must be assigned to whoever can change the system's behavior, not to whoever happens to operate it and not to a committee.

The second shift is that disclaimers have twice failed as a defense. A tribunal refused to accept that a customer should have clicked through to the authoritative page; a court refused to accept that a user's theoretical ability to cross-check discharged the duty. Warnings are still worth publishing, for honesty. They aren't worth relying on.

CLAUSES

1.1 **SHALL** Every AI system in the inventory required by Article 4 have a single named Accountable Owner who is an individual, not a role, a team, a committee or a vendor. The name SHALL appear in the inventory record and SHALL be re-attested at least annually and on any change of ownership.

1.2 **SHALL** The Accountable Owner have the authority to suspend, restrict or decommission the system without requiring approval from the function that operates it commercially.

1.3 **SHALL** Every AI system have a documented escalation path naming, for each of at least 3 severity levels, who is notified, within what time, and who holds the authority to stop the system.

1.4 **SHALL** The organization treat AI-generated output published to, or acted upon in respect of, a third party as a statement of the organization, of the same character as a statement by an employee. No policy, contract or interface SHALL characterize the AI system as a separate actor, a distinct legal entity, or as being responsible for its own statements.

1.5 **SHALL NOT** The organization rely on a disclaimer, a hyperlink to authoritative content, or a general statement that outputs may be inaccurate as its principal control against incorrect output reaching a third party. Disclaimers are a transparency measure under Article 11, not a correctness control under Article 12.

1.6 **SHALL** Accountability for a system that makes or materially influences a decision about a natural person be assigned at a level of seniority commensurate with the consequence of that decision and SHALL be recorded in the minutes of the body that approved deployment.

1.7 **SHOULD** The organization maintain a standing governance body with cross-functional membership that at minimum includes legal, security, privacy, records management and the business function, and SHOULD include human resources where the system touches employment. Governance composed only of technologists reproduces the failure pattern the 2026 survey evidence describes.

1.8 **SHALL** Where a supplier operates an AI system on the organization's behalf, the contract identify the supplier's Accountable Owner by name and require notification within a defined period of any change.

EVIDENCE

Evidence artifact	Contents / test of adequacy	Owner
AI Accountability Register	1 row per system: system identifier, Accountable Owner (named individual), approving body, date of last attestation, suspension authority. Test: pick any production AI system at random; the owner must be reachable and must confirm the assignment.	Governance
Escalation and stop procedure	3 severity tiers, named notification targets, maximum notification times, named stop authority, and the technical mechanism referenced (see Article 8).	Operations
Deployment approval minutes	For each system touching natural persons: record of the approving body, date, accountable owner and the risk classification relied upon.	Governance
Contractual owner clauses	Executed supplier agreements naming the supplier-side accountable individual and change-notification duty.	Legal

Article 2 — Declared purpose, non-goals and proportionate value

THE PRINCIPLE

An AI system is the result of intentional design, not glue code around a powerful API. We don't ship because we can. We ship when we understand why we should, and we say in advance what the system is not for.

WHY IT CHANGED

V1 required a purpose statement and explicit non-goals. It did not require anyone to say what value the system was supposed to produce, how that would be measured, or what would cause it to be switched off. The 2026 evidence base shows this omission is the root cause of ungovernable estates: organizations accumulate pilots that cannot be killed because nobody ever agreed what success looked like, and most executives now identify the inability to connect AI to profit and loss as their principal barrier.

A decommissioning trigger agreed at approval is the cheapest governance control in this document. It converts "should we stop this?" from a political question into an arithmetic one.

CLAUSES

2.1 **SHALL** Every AI system have an approved Purpose Statement of no more than 1 page, recorded before development begins, stating: the problem, the population affected, the decision or task the system participates in, and the human process it changes.

2.2 **SHALL** Every Purpose Statement includes explicit non-goals (the uses the system must not be put to), and these SHALL be enforced technically wherever technically possible, not only stated.

2.3 **SHALL** Every AI system has a declared Value Hypothesis at approval stating the expected benefit, the metric by which it will be measured, the baseline against which it will be compared, and the date of first measurement.

2.4 **SHALL NOT** Self-reported user perception of productivity or time saved be used as the sole evidence for a Value Hypothesis. Controlled evidence indicates perceived and actual productivity effects diverge substantially and in both directions.

2.5 **SHALL** Every AI system has a Decommissioning Trigger agreed at approval: a stated condition of value, accuracy, cost or risk which, if met, requires the Accountable Owner to formally re-approve, restrict or retire the system.

2.6 **SHALL** The organization review every AI system against its Value Hypothesis and Decommissioning Trigger at least annually, and record the outcome as continue, restrict, or retire.

2.7 **SHOULD** The organization prefer narrowing a system's scope over deploying it at a reliability level below its intended use. Where reliability cannot be achieved for the declared purpose, the correct outcome is a narrower purpose or no deployment.

2.8 **SHALL** Any external claim about the organization's AI capability (in marketing, investor communications, procurement responses or regulatory filings) be traceable to a system in the inventory and to evidence of the capability claimed. Overstatement of AI capability is a disclosure offence in multiple jurisdictions and has resulted in both civil penalties and criminal charges.

EVIDENCE

Evidence artifact	Contents / test of adequacy	Owner
Purpose Statement (per system)	1 page: problem, affected population, task, human process changed, explicit non-goals, technical enforcement of non-goals.	Product
Value Hypothesis and measurement record	Metric, baseline, measurement date, measured result, and the method — which must not be self-report alone.	Product / Finance
Decommissioning Trigger and annual review	Stated trigger condition; dated annual decision recorded as continue, restrict or retire, with signature.	Governance
AI capability claim register	Every external AI claim, the system it refers to, and the evidence supporting it. Reviewed before publication.	Marketing / Legal

Article 3 — Human dignity and differential impact

THE PRINCIPLE

If a shortcut is unethical, it isn't a shortcut. It's a liability. We actively seek and mitigate bias rather than hoping it is small enough. We avoid dark patterns, manipulative interfaces and deceptive personification. We consider how different communities are affected differently, and we measure success by human outcomes, not only business metrics.

WHY IT CHANGED

This Article merges 2 V1 principles, ethics over efficiency and inclusion, because in practice they are assessed together and because keeping them apart let both remain unmeasured. The substance hasn't changed. What's new is that the aspiration now has a method (AI system impact assessment is a published international standard), a legal edge in employment (vendor-as-agent liability, a certified nationwide collective, state statutes reaching unintentional discrimination), and a documented anti-pattern: an obligation that triggers only on self-classification will be defined around.

The other addition is deceptive personification, which V1 named and which has since become a distinct regulatory category. 14 United States states enacted companion or chatbot safety legislation in 2026 alone, and EU transparency obligations require that a person be told they are interacting with an AI system.

CLAUSES

3.1 **SHALL** Every AI system that participates in a decision affecting a natural person's employment, credit, housing, education, healthcare, insurance, access to essential services or legal standing undergo a documented Impact Assessment before deployment and on material change, following a recognized method.

3.2 **SHALL** The Impact Assessment trigger on the system's function, not on the organization's classification of the system. Whether the organization calls the system an automated decision tool is irrelevant; whether it materially influences such a decision is decisive.

3.3 **SHALL** Differential performance testing be conducted across the populations the system affects, with results recorded numerically, retained, and compared against the previous test at each review.

3.4 **SHALL** Where differential performance is identified and cannot be mitigated, the system be restricted or withdrawn from the affected use, and the decision recorded. Documenting a disparity without acting on it is worse than not testing, because it establishes knowledge.

3.5 **SHALL NOT** An AI system present itself as a human being, adopt a persona designed to induce misplaced trust, or withhold its artificial nature from a person who is interacting with it.

3.6 **SHALL NOT** An interface use AI-generated content, urgency, personalization or simulated relationship to induce a decision the person would not otherwise make. Persuasion by accurate information is legitimate; manufactured intimacy is not.

3.7 **SHALL** A person subject to a materially automated decision be able to obtain a plain-language explanation of the principal factors, and to reach a human with authority to review the decision. This is an operative legal right in several jurisdictions and SHOULD be implemented uniformly rather than per-jurisdiction.

3.8 **SHOULD** Affected groups, domain experts and front-line staff be consulted in design and evaluation, and the consultation recorded. A system evaluated only by the people who built it has been evaluated by its most optimistic possible reviewers.

3.9 **SHALL** The organization consider the effect of AI deployment on the quality of work received by colleagues. Output that appears complete but shifts verification burden onto others is a governance failure with a measurable organizational cost, not a personal style choice.

EVIDENCE

Evidence artifact	Contents / test of adequacy	Owner
AI Impact Assessment (per qualifying system)	Affected populations; harms considered; differential performance results with numbers; mitigations; residual risk; approval signature; review date.	Governance / Legal
Differential performance test record	Metric by population segment, dated, with comparison to prior test. Retained for the life of the system plus the applicable limitation period.	Engineering
Human review procedure	Documented route by which an affected person reaches a human with authority to alter the decision, with service levels.	Operations
Consultation record	Who was consulted, when, what changed as a result.	Product

Article 4 — Inventory and Bill of Materials

THE PRINCIPLE

We cannot govern what we have not enumerated. Every AI system, model, agent and integration the organization operates or embeds is registered, versioned, owned and traceable to its components.

WHY IT CHANGED

This Article didn't exist in V1, which is its own indictment: every other principle in V1 is unenforceable without it. The 2026 evidence indicates most organizations can't satisfy it. Survey work reports that a large majority have AI agents in their environments that nobody registered, and that many discovered unknown agents repeatedly over a single year.

This is also among the first questions enterprise assurance reviews now ask after governance, and they ask it in a revealing way: not "do you have an inventory" but "is it centralized and automated, or is it a spreadsheet somebody updates periodically?" That distinction is the difference between a Low and a Medium risk rating, and it's a fair one. A manual inventory is a snapshot of what people remembered to declare.

The bill of materials requirement is newer still. Machine-readable formats capable of expressing model and dataset composition now exist, and in May 2026 an international minimum-elements specification for AI bills of materials was published jointly by the United States cyber-defense agency and G7 partners, organized into 7 clusters: metadata, system-level properties, models, dataset properties, infrastructure, security properties and key performance indicators. Its authors declined to include a standalone element for level of autonomy, so Article 8 requires it separately.

CLAUSES

4.1 **SHALL** The organization maintain a single authoritative AI Inventory covering every AI system it develops, operates, embeds, or consumes as a service, including AI features enabled inside third-party products, whether or not the organization chose to enable them.

4.2 **SHALL** Each inventory record contain at minimum: unique identifier; purpose statement reference; Accountable Owner; risk classification; the models and model versions in use; the data domains accessed; the tools and systems the AI may invoke; the retention window applying to its inputs and outputs; the hosting location and jurisdiction; the sub-processors involved; and the date of last review.

4.3 **SHALL** Registration be a precondition of production deployment, enforced by a technical control in the deployment pipeline rather than by a procedural instruction.

4.4 **SHOULD** The inventory be populated by automated discovery from identity, network, cloud, code and expense telemetry in addition to self-declaration. A registry that depends solely on people declaring what they built will systematically omit whatever was built quickly.

4.5 **SHALL** The organization conduct periodic discovery for unregistered AI use (including personal-account use of public AI services on organizational data) and record findings. Telemetry-based research in February 2026 indicated that most interactions with several leading assistants occurred through unmanaged personal accounts, which are outside the reach of both enterprise agreements and data-loss controls.

4.6 **SHALL** The organization be able to produce, on request and within a defined period, an AI Bill of Materials for any registered system, expressed in a machine-readable format, covering: model identity and version; provenance and lineage including base model and any fine-tuning; training,

validation and test dataset references with provenance and sensitivity classification; framework and runtime dependencies; cryptographic hashes of weights where the organization controls them; licences for both model and data; and security properties.

4.7 **SHALL** The AI Bill of Materials record the system's level of autonomy, notwithstanding that current minimum-element specifications omit it.

4.8 **SHALL** Decommissioned systems remain in the inventory, marked as retired, with the date and the disposition of their data and logs. An inventory that deletes history cannot answer a discovery request.

EVIDENCE

Evidence artifact	Contents / test of adequacy	Owner
AI Inventory	Complete register per clause 4.2. Test of adequacy: reconcile against cloud spend, identity grants and code repositories; unexplained gaps are the finding.	Governance / IT
Deployment gate control	Evidence that unregistered systems cannot reach production — pipeline configuration, not policy text.	Engineering
Shadow AI discovery report	Periodic scan results, unregistered use identified, remediation and dates.	Security
AI Bill of Materials (per system)	Machine-readable, generated from build rather than hand-authored, including autonomy level.	Engineering
Retired systems record	Retirement date, data disposition, log disposition, hold status.	Records

Article 5 — Data provenance and rights

THE PRINCIPLE

We know where our models came from, what they were trained on, and on what legal basis. We don't use data we don't have the right to use, and we don't turn our customers' data into someone else's model.

WHY IT CHANGED

V1 addressed data protection but never model provenance. That omission is now the largest source of commercial exposure in the field, and the shape of the exposure is counter-intuitive: the largest settlement on record was paid by a party that had substantially won the fair-use argument, because the copies underpinning its library had been acquired unlawfully. Acquisition legality, not use doctrine, is what determined liability.

Meanwhile, buyer scrutiny has become specific. Enterprise buyers now ask, in sequence, whether the model is used off-the-shelf or retrieval-augmented or fine-tuned; what documentation exists on training data; whether other clients' data was used; whether personal data was used; and how the buyer's data is isolated from other clients' training sets. A supplier who has not decided its answers to those 5 questions in advance will answer them inconsistently under pressure.

CLAUSES

5.1 **SHALL** The organization record, for every model in use, its origin, its license, and the mode in which it is used (off-the-shelf inference, retrieval-augmented, fine-tuned, or trained by the organization) and SHALL be able to state that mode without qualification.

5.2 **SHALL** The organization hold, or be able to obtain from the upstream provider, documentation of the training data, methodology and known limitations of each model in use, in the form of a model card, factsheet or transparency report.

5.3 **SHALL NOT** Customer or client data be used to train, fine-tune or otherwise update model weights without a separate, explicit, written authorization from that customer that is not bundled into general terms of service.

5.4 **SHALL** Where the organization trains or fine-tunes on any data, it records the legal basis for each dataset, the method of acquisition, and evidence that acquisition was lawful. Provenance of acquisition SHALL be documented to the same standard as provenance of use.

5.5 **SHALL** Where customer data is used in a permitted training or tuning process, that data be isolated such that it cannot influence a model serving any other customer, and the isolation SHALL be technically enforced and describable.

5.6 **SHALL** The organization distinguish, in every representation it makes, between model training and data processing. Personnel accessing production personal data for system testing is a processing activity governed by the data protection agreement; it is not model training. Conflating the two in a questionnaire response creates a misrepresentation risk in both directions.

5.7 **SHALL** Where personal data is present in any dataset used for training, validation or testing, the legal basis be documented, a privacy impact assessment completed, and the record retained.

5.8 **SHOULD** The organization prefer hosted model access over the distribution of downloadable model weights where the intellectual-property exposure profile is a material consideration, in light of

judicial reasoning that model weights are not themselves an infringing copy but distribution changes the analysis.

5.9 **SHALL** Contracts with model and platform providers be reviewed at least annually for changes to training, retention and data-use terms, and the review recorded. A supplier can change the AI data posture of the organization's entire estate through a terms update.

EVIDENCE

Evidence artifact	Contents / test of adequacy	Owner
Model provenance register	Per model: origin, version, licence, usage mode, model card location, date of last terms review.	Engineering / Legal
Training data legal basis file	Per dataset: source, acquisition method, evidence of lawful acquisition, licence, personal-data assessment.	Legal
Isolation architecture description	Technical description of how customer data is prevented from influencing models serving others, with the control that enforces it.	Engineering
Terms change review log	Dated annual review of each provider's training and retention terms, with material changes escalated.	Legal / Procurement

Article 6 — Custody, retention and records

THE PRINCIPLE

Prompts and outputs are records. They are evidence of how we reasoned, what we told people, and what we relied upon. We decide deliberately how long we keep them, who may read them, and how we produce them when they are demanded. We decide before anyone demands them.

WHY IT CHANGED

This Article has no ancestor in V1, which mentioned retention only in the phrase "for exactly as long as it's needed." That formulation is now actively dangerous, because it presumes the organization is choosing. In practice retention is being set by third-party platform defaults of 90 to 180 days, and in at least one major matter the assumed policy and the actual configuration differed by a factor of 4.

The legal ground has moved decisively. A United States court has ordered production of the specific prompts an expert used, holding that methodology is fair ground for discovery. Another has ordered production of some 20 million de-identified assistant conversation logs. A financial regulator requires AI chatbot communications to be captured in books and records. And the leading eDiscovery standards body did not begin drafting guidance on any of this until July 2026, with publication targeted for April 2027.

The likeliest failure isn't a bad decision. It's no decision. Evidence is destroyed automatically, before anyone knows a hold applies, by a configuration nobody chose. Recording the actual retention window per platform (the configured value, verified, not the assumed one) is the highest-value control in this document relative to its cost.

CLAUSES

6.1 **SHALL** The organization determines and records, for every AI system in the inventory, whether its inputs and outputs constitute business records, and under which retention schedule they fall.

6.2 **SHALL** The actual configured retention period for prompts, outputs and logs be recorded per platform in the inventory and SHALL be verified against the platform configuration rather than assumed from contract or memory. The verification date SHALL be recorded.

6.3 **SHALL** The organization be able to suspend automatic deletion of prompts, outputs and logs for any AI system within a defined period of a legal hold being issued, and SHALL have tested that capability.

6.4 **SHALL** The legal hold procedure explicitly enumerates AI artifacts (prompts, outputs, system instructions, tool-call records, retrieval context, model and prompt versions, and evaluation results) and name each platform on which they reside, including personal-account and local-device locations.

6.5 **SHALL** The organization state, without ambiguity, whether it retains customer inference inputs and outputs at rest; where derived AI outputs materialized into a customer's own environment by design, that SHALL be described explicitly rather than characterized as zero retention.

6.6 **SHALL** Human access by the organization's personnel or contractors to the content of customer inputs and outputs be defined, minimized, logged, and disclosed. Where personnel access live production personal data as part of delivery, that SHALL be stated plainly and SHALL match the data protection agreement.

6.7 **SHALL** Content logging for troubleshooting be subject to a defined retention period, with sensitive data redacted at the point of ingestion where the log is retained beyond the immediate operational need.

6.8 **SHALL** Where the organization is contractually or legally permitted to use inputs for training, the retention and disposal policy for those raw datasets be stated, with a defined disposal method and timeline.

6.9 **SHALL** Personnel be instructed, in an enforceable acceptable use policy, not to input privileged or legally sensitive material into AI platforms the organization does not control, and the instruction SHALL be reissued when new tools are adopted.

6.10 **SHOULD** The organization conduct a periodic AI usage audit identifying every platform in use, its retention period, its export capability and its preservation steps, and SHOULD treat the absence of an export capability as a procurement defect.

EVIDENCE

Evidence artifact	Contents / test of adequacy	Owner
Retention schedule for AI artifacts	Per platform: artifact types, configured retention, verification date, records classification, disposal method.	Records / Legal
Legal hold procedure — AI annex	Enumerated artifact types; platform-by-platform preservation steps; suspension mechanism; tested date and result.	Legal
Human access policy and access log	Who may read customer content, under what conditions, with what approval, and the log demonstrating it.	Security / Legal
Redaction control evidence	Configuration demonstrating redaction at ingestion for retained content logs.	Engineering
Periodic AI usage audit	Platform list, retention windows, export capability, gaps and remediation.	Records / Security

Article 7 — Supply chain and sub-processors

THE PRINCIPLE

The obligations we accept, we impose. Every party that touches our customers' data through our AI is named, contractually bound, and technically verified. We do not accept "to improve our services" as a description of what happens to that data.

WHY IT CHANGED

V1 governs what the organization builds and says nothing about what sits underneath it. That is exactly where the customer's question lands: enterprise questionnaires ask suppliers to list all third-party AI sub-processors and confirm they are contractually prohibited from using the customer's data for their own model training. They also distinguish between having a contract clause and having verified that the training feature is actually disabled.

The evidence justifies the skepticism. Analysis of some 2,400 software providers found that nearly 64 percent failed to disclose sub-processing conducted by another AI provider. Separately, an AI chat integration holding an OAuth grant was used in August 2025 to reach data in more than 700 organizations without triggering multi-factor authentication. The AI was never compromised. Its authorization was.

THE SENTENCE TO INTERNALISE

An AI integration is a privileged, often MFA-bypassing, non-human identity inside your systems, frequently granted by a business user, and it persists after the person who granted it leaves.

CLAUSES

7.1 **SHALL** The organization maintain a register of every third-party AI sub-processor, identifying the service, the data categories it receives, the processing location, and the contractual basis on which it may not use that data for model training.

7.2 **SHALL** The register extend to fourth parties (the sub-processors of the organization's sub-processors) to the extent obtainable and SHALL record where that information could not be obtained.

7.3 **SHALL NOT** The organization accept contractual language permitting use of its data or its customers' data to "improve the service" without a specific, written carve-out excluding model training and a defined opt-out mechanism.

7.4 **SHOULD** The organization verify technically, as well as contractually, that training and retention features are disabled (by administrative console configuration, API response, or written provider attestation specific to the organization's tenant), and SHOULD record the verification date and method.

7.5 **SHALL** Every AI integration authorized by OAuth, API key or service credential be inventoried as a non-human identity, scoped to least privilege, owned by a named individual, reviewed at least quarterly, and revocable within a defined period.

7.6 **SHALL** The organization monitor supplier terms for changes to AI data-use defaults and SHALL treat the enabling of an AI feature that processes organizational data by default as a change requiring assessment, not as a product update.

7.7 **SHALL** Supplier agreements require notification of AI-related security incidents within a defined period and SHALL require the supplier to identify the AI components involved.

7.8 **SHOULD** The organization require, from suppliers of material AI capability, either an independent assurance artifact under Article 13 or a completed AI-specific assessment questionnaire and SHOULD record which was accepted and why.

7.9 **SHALL** Where the organization is itself the supplier, it be able to produce its sub-processor register to a customer on request and SHALL notify customers of material changes before they take effect.

EVIDENCE

Evidence artifact	Contents / test of adequacy	Owner
AI sub-processor register	Service, role, data categories, location, no-training basis, verification method and date, fourth parties known and unknown.	Procurement / Legal
Non-human identity inventory	Every AI integration credential: scope, owner, grant date, last review, revocation path and tested revocation time.	Security
Technical verification records	Console screenshots, API responses or tenant-specific attestations evidencing training and retention settings.	Security
Supplier terms monitoring log	Dated record of AI-related terms changes, assessment outcome, action taken.	Procurement
Customer-facing sub-processor disclosure	The published or contractually delivered list, with change-notification history.	Legal

Article 8 — Bounded agency

THE PRINCIPLE

An AI system may act only within limits a human set in advance, using credentials a human granted for that purpose, and it can always be stopped. Capability is not permission.

WHY IT CHANGED

V1 has no concept of an AI system that acts. It reasons entirely about systems that produce outputs which humans then use. Every serious failure of the past 2 years lives in the gap that assumption creates. The 2026 threat taxonomies now place excessive agency in the top 3 risks for language-model applications and devote an entire companion list to agentic applications, introducing the principle of least agency alongside least privilege.

Two lessons need stating outright. First, an agent that violates an explicit freeze and destroys production data isn't hypothetical. It happened in July 2025, to a company that builds developer tools. Second, and less discussed: in that incident the agent then reported that recovery was impossible, which was false. An agent's account of what it did is generated text. It is not telemetry.

Never accept an agent's self-report as evidence of system state.
Verification comes from the system (logs, database state, API responses), not from asking the agent what happened. This is a design requirement, not an operational preference.

CLAUSES

8.1 **SHALL** Every AI system be assigned a declared Autonomy Level, recorded in the inventory and the bill of materials, drawn from a defined scale: (0) advisory only, no action; (1) acts with per-action human approval; (2) acts within a pre-approved allowlist and reports; (3) acts autonomously within a bounded domain with post-hoc review.

8.2 **SHALL** Every tool, API, dataset and system an AI agent may invoke be enumerated explicitly. Agents SHALL NOT be granted broad or wildcard scopes. Enumerated permission is the control surface; egress inspection is not, because an agent exfiltrating data through an authorized integration will not trigger data-loss prevention.

8.3 **SHALL** Mandatory human approval be required before an AI system takes any action that is externally visible, financially material, legally binding, destructive, or irreversible, including sending external communications, deleting or modifying records of record, transferring funds, changing access rights, or publishing content.

8.4 **SHALL** Every AI agent operate under a distinct, attributable identity (not a shared service account and not a human's credentials) so that its actions can be isolated in logs and its access revoked without affecting anyone else.

8.5 **SHALL** The organization maintain a documented emergency shutdown procedure capable of disabling an AI system or agent, which SHALL specify: the trigger conditions; who may invoke it without further approval; the technical steps in order; the expected time to full effect; how credential and token revocation is achieved; how in-flight actions are handled; and who is notified. The procedure SHALL be tested at least annually and the test recorded.

8.6 **SHALL** Agents operating on production systems be technically prevented from acting on production data outside a declared scope, with environment separation enforced rather than instructed.

8.7 **SHALL** Where an agent may generate and execute code, execution occur in an isolated environment (at minimum containerized with non-root privileges, restricted namespaces, resource limits and network egress filtering), and the organization SHALL state whether generated code is executed at runtime at all, since many systems generate code only in a development pipeline under human review.

8.8 **SHALL** Inter-agent and agent-to-tool communication be authenticated, and tokens SHALL NOT be passed through to downstream services that did not issue them.

8.9 **SHALL** Third-party agent tools and connectors be treated as untrusted supply chain components, assessed before adoption and version-pinned, considering documented tool-poisoning and rug-pull techniques now catalogued in adversarial knowledge bases.

8.10 **SHALL** The organization record every agent action in an append-only log sufficient to reconstruct what was done, by which agent identity, under whose authority, with what inputs, and with what effect, independent of anything the agent itself reports.

EVIDENCE

Evidence artifact	Contents / test of adequacy	Owner
Autonomy Level declaration	Per system, on the defined 0–3 scale, with justification and approval.	Product / Governance
Tool permission manifest	Enumerated tools, APIs and datasets per agent, with scope, justification and quarterly re-attestation of minimality.	Engineering / Security
Emergency shutdown procedure	Triggers, authority, ordered steps, time to effect, credential revocation, in-flight handling, notification list; annual test record.	Operations
Agent identity register	Distinct identity per agent, owner, scope, issue and review dates, revocation path.	Security
Agent action log	Append-only, tamper-evident, retained per Article 6, independent of agent self-report.	Engineering
Human approval gate evidence	Configuration demonstrating that high-consequence action classes cannot execute without approval.	Engineering

Article 9 — Security by design

THE PRINCIPLE

Security is part of safety. There is no AI Done Right without secure handling of prompts, data, secrets and outputs; defense against injection, exfiltration and abuse; least privilege everywhere; and an incident response plan that anticipates the model being manipulated rather than assuming it won't be.

WHY IT CHANGED

V1 named prompt injection and data exfiltration, which was prescient. What V1 assumed, and what 2026 disproves, is that these are preventable. The chief information security officer of a leading model provider stated publicly in October 2025 that prompt injection remains a frontier, unsolved security problem, and his company conceded in December 2025 that it may never be fully solved. The 2026 threat list's own framing is to build the system around the model so that when the model is fooled, nothing important breaks.

That's the correct architecture, and it's a different instruction from "prevent injection." It means the security objective is blast-radius containment, and it is achieved through the controls in Article 8 far more than through input filtering.

CLAUSES

9.1 **SHALL** The organization design on the assumption that the model's instruction boundary can be crossed and SHALL demonstrate that a successful injection cannot by itself cause a material loss. Controls that depend on the model refusing SHALL NOT be the sole control for any high-consequence action.

9.2 **SHALL** System instructions and privileged context be isolated from user-supplied and retrieved content by an architectural boundary, and the organization SHALL treat all retrieved content (documents, web pages, tool results, form fields, file metadata) as untrusted input.

9.3 **SHALL** Baseline enterprise controls apply to all AI components without exception: encryption in transit and at rest, multi-factor authentication, role-based access control, secrets management, and network segmentation.

9.4 **SHOULD** An independent guardrail layer inspect inputs and outputs, and outbound network destinations reachable by AI components SHOULD be allowlisted. Where an organization cannot yet meet this, it SHALL record the gap and the target date rather than describing its prompt engineering as equivalent.

9.5 **SHALL** The organization conduct adversarial testing of AI systems before each material release, covering at minimum: prompt injection including indirect and cross-modal vectors; system-instruction and hidden-context extraction; unsafe output handling; tool misuse; and unbounded consumption. Results SHALL be recorded and retained.

9.6 **SHALL** Incident response procedures explicitly cover AI-specific incidents (manipulated output, agent misbehavior, data exfiltration through an authorized integration, model or tool supply-chain compromise) with named severity criteria and regulatory reporting timelines where applicable.

9.7 **SHALL** Vulnerability management extend to models, agent frameworks, connectors and vector stores, with a defined path for applying provider-side fixes and for withdrawing a component that cannot be fixed.

9.8 **SHOULD** The organization map its AI threat model to a recognized adversarial taxonomy so that findings are expressed in vocabulary an assessor already understands.

9.9 **SHALL** Where AI systems process personal data, retrieval and tool access be scoped to the requesting user's existing entitlements, so that the AI cannot become a mechanism for over-broad access to data the user could not otherwise reach.

EVIDENCE

Evidence artifact	Contents / test of adequacy	Owner
AI threat model	Per system: assets, adversaries, attack paths mapped to a recognized taxonomy, controls and residual risk.	Security
Adversarial test report	Scope, vectors tested, findings, severity, remediation status, date, tester identity.	Security
Guardrail and egress configuration	Evidence of input/output inspection and outbound allowlisting, or a recorded gap with target date.	Engineering
AI incident response annex	AI-specific scenarios, severity criteria, notification timelines, tested date.	Security
Entitlement scoping evidence	Demonstration that retrieval respects the requesting user's permissions.	Engineering

Article 10 — Resource and cost containment

THE PRINCIPLE

An AI system that can spend without limit is an availability risk and a financial one. Consumption is bounded, attributed and observable.

WHY IT CHANGED

V1 treats scalability as technical resilience and governance capacity and never mentions money. In the 2026 threat taxonomy, unbounded consumption rose 4 places into the upper half of the risk list, because it is simultaneously a denial-of-service vector, a cost-exhaustion attack and, in agentic systems that can invoke themselves, an unbounded loop with a billing meter attached. Enterprise questionnaires now ask about quotas and automated circuit breakers directly and distinguish between hard technical limits and after-the-fact monitoring.

There's a second reason for this Article. Cost per outcome is the denominator of the value hypothesis required by Article 2. An organization that cannot attribute AI spend to a system cannot evaluate whether that system should continue.

CLAUSES

10.1 **SHALL** Every AI system operate under enforced consumption quotas at more than one level (at minimum per identity and per system), expressed in tokens, requests or cost.

10.2 **SHALL** Automated circuit breakers halt or throttle consumption when a defined volumetric or financial threshold is exceeded, without requiring human intervention to take effect.

10.3 **SHALL** Agentic systems have a bounded maximum number of steps, tool invocations and recursive self-invocations per task, enforced technically.

10.4 **SHALL** AI consumption be attributable to a system in the inventory and, where applicable, to a customer, so that cost per outcome can be computed.

10.5 **SHOULD** Where the organization supplies AI capability to a customer under a fixed price, the allocation of consumption risk be stated explicitly in the contract, so that neither party discovers it during an incident.

10.6 **SHALL** Consumption anomalies generate an alert to a named owner within a defined period, and the organization SHALL retain evidence of alert configuration.

10.7 **SHOULD** Materialized AI outputs be reused rather than regenerated where the input is unchanged, both to contain cost and to make results reproducible for audit.

EVIDENCE

Evidence artifact	Contents / test of adequacy	Owner
Quota and circuit breaker configuration	Per-identity and per-system limits, thresholds, automatic action taken, evidence from configuration.	Engineering / FinOps
Agent step and recursion bounds	Configured maxima per agent task type.	Engineering
Cost attribution report	AI spend mapped to inventory identifiers and, where applicable, to customers.	Finance
Consumption alerting evidence	Alert rules, named owners, response times, sample incident.	Operations

Article 11 — Transparency and disclosure

THE PRINCIPLE

AI shouldn't be a sealed box that people are simply asked to trust. We explain in plain language what the system does, what data it uses and how outputs are generated; we mark what we generate; and we give people enough understanding to make informed choices. Transparency is not total source exposure.

WHY IT CHANGED

V1 framed transparency as an ethical courtesy. It became a legal obligation on August 2nd, 2026, when the European transparency regime became applicable. It requires that people be told they are interacting with an AI system, and that synthetic content be marked in a machine-readable format that is effective, interoperable, robust and reliable as far as technically feasible. A voluntary Code of Practice finalized in June 2026 asks signatories to apply at least 2 layers of machine-readable marking, chosen from metadata, watermarking or other measures, and states expressly that no single technique suffices. A Californian statute imposing latent and manifest disclosure and a public detection tool on large generative providers became operative on the same date, with platform duties following in 2027 and capture-device duties in 2028.

The relevant technical specification is a content provenance standard now at version 2.4 with a conformance program and a trust list; the corresponding international standardization work is in progress but not yet published. Organizations should implement the specification and avoid claiming it is an international standard.

CLAUSES

11.1 **SHALL** A person interacting with an AI system be informed of that fact at the point of interaction, in plain language, without requiring them to seek the information out.

11.2 **SHALL** Synthetic audio, image and video content the organization generates and publishes be marked in a machine-readable format conveying its artificial origin, using a recognized content provenance specification.

11.3 **SHALL** AI-generated or AI-manipulated content published on matters of public interest be labelled visibly to the recipient, in addition to machine-readable marking.

11.4 **SHALL NOT** The organization strip, degrade or fail to propagate provenance metadata present in content it processes, where preservation is technically feasible.

11.5 **SHALL** The organization maintain a system card or equivalent for each material AI system, describing purpose, capabilities, limitations, known failure modes, data used, human oversight arrangements and evaluation results, at a level of detail it is prepared to share with customers and regulators on request.

11.6 **SHALL** Limitations and known failure modes be disclosed proactively to users of the system, not only on request. Honesty about uncertainty is a control, not a marketing liability.

11.7 **SHOULD** Where an AI-augmented value is presented to a user, the interface surface its provenance and a confidence indication and permit the user to see what the value was derived from. Where the organization already materializes provenance and confidence for internal purposes, surfacing it is a small increment that materially changes the transparency posture.

11.8 **SHALL** Disclosures be consistent across the questionnaire responses, contractual representations, privacy notices and public statements the organization makes. Inconsistency between these is itself a finding, and in regulated sectors an enforcement risk.

EVIDENCE

Evidence artifact	Contents / test of adequacy	Owner
AI interaction disclosure	Screenshots or interface specification demonstrating point-of-interaction disclosure.	Product
Provenance marking implementation	Specification and version implemented, signing arrangement, verification method, coverage by content type.	Engineering
System card (per material system)	Purpose, capabilities, limitations, failure modes, data, oversight, evaluation results, version and date.	Product
Disclosure consistency review	Dated cross-check of questionnaire answers, contracts, privacy notices and marketing claims.	Legal

Article 12 — Evaluation, monitoring and drift

THE PRINCIPLE

Powerful models aren't a substitute for engineering rigor. We test the model's behavior in realistic scenarios as well as the code around it; we monitor continuously for accuracy, bias, abuse, drift and security; and when we learn, we adapt. Deployment is the beginning of responsibility.

WHY IT CHANGED

This Article is V1's 2 strongest principles fused and given a number. It needs a number because of the most damning control statistic of 2026: 84 percent of technology leaders report having no formal process to track whether their AI is accurate. Every other failure in Part One is downstream of that gap: the fabricated citations, the refunded consulting report, the chatbot misinformation, the composed falsehood in a search summary.

A second change is that "drift" now has 2 meanings. Model drift in the classical sense (performance degrading as the world moves) still happens. But for organizations consuming hosted models, the more common and less monitored event is that the provider changed the model underneath them. Version pinning and change detection are therefore control requirements, not engineering preferences.

CLAUSES

12.1 **SHALL** Every AI system have a defined accuracy or quality measure appropriate to its purpose, a stated acceptable threshold, and a documented method of measurement agreed before deployment.

12.2 **SHALL** Accuracy be measured on a defined schedule after deployment, the results recorded and retained, and a breach of threshold SHALL trigger the escalation path in Article 1.

12.3 **SHALL** The organization maintain an evaluation set representative of real use, including known-hard and adversarial cases, and SHALL re-run it before each material change to model, prompt, retrieval corpus or tooling.

12.4 **SHALL** Model and prompt versions be pinned, recorded with every material output, and changed deliberately. Provider-initiated model changes SHALL be detected and treated as a change requiring re-evaluation.

12.5 **SHALL** Where AI output is used in any document or communication that asserts facts, citations or figures to a third party, a verification step confirms those assertions before release, and evidence of verification SHALL be retained. This clause exists because of documented professional-services failures and judicial sanctions arising from unverified generated citations.

12.6 **SHALL** AI-augmented data transformations record provenance, a confidence indication, and the model and prompt version, and results SHALL be materialized such that they are reproducible without re-running inference.

12.7 **SHALL** Monitoring cover, at minimum: accuracy against threshold; differential performance per Article 3; abuse and jailbreak attempts; consumption anomalies per Article 10; and provider-side model change.

12.8 **SHALL** The organization maintain a feedback channel for users and affected persons to report incorrect or harmful output, with a defined triage and response process and evidence of follow-through.

12.9 **SHOULD** The organization publish or make available on request a change log describing what changed in the system, why, and with what measured effect.

12.10 **SHALL** Where reliability cannot be brought to the level the declared purpose requires, the purpose be narrowed or the system withdrawn, per Article 2.

EVIDENCE

Evidence artifact	Contents / test of adequacy	Owner
Accuracy measurement record	Metric, threshold, schedule, measured results over time, threshold breaches and actions taken.	Engineering / Product
Evaluation set and results	Versioned evaluation corpus including adversarial cases; results per release.	Engineering
Version pinning and change detection	Model and prompt versions recorded per output; provider change alerts and re-evaluation records.	Engineering
Fact verification evidence	For external documents: who verified which assertions, when, and against what source.	Delivery / Quality
Feedback and remediation log	Reports received, triage, outcome, time to resolution.	Operations

Article 13 — Independent assurance and evidence

THE PRINCIPLE

We are prepared to be checked. Conformity that only we can verify is not conformity; it is confidence. We maintain the evidence, and we submit it to people who do not work for us.

WHY IT CHANGED

V1 ends with a pledge, and a pledge cannot be failed. This Article exists because the market changed underneath that position. Accredited certification against an AI management system standard is now commercially available; a security assurance registry makes that certificate a precondition of its higher tier; a certifiable AI security assessment exists; an agentic standard has accredited auditors. So "no independent testing has been performed" has moved from ordinary to conspicuous, and enterprise questionnaires ask the question in a way that makes internal audit visibly the third-best answer of 4.

This is also the Article on which most organizations, well-run ones included, will score worst at first. That's no reason to soften it. It's a reason to plan the path deliberately and to say so honestly in the interim, which is defensible in itself and far better than an unsupported claim that collapses under a document request.

CLAUSES

13.1 **SHALL** The organization maintain an Evidence Register listing, for every artifact required by this standard, its location, its owner, the date it was last refreshed, and the review cycle.

13.2 **SHALL** The organization be able to produce any artifact in the Evidence Register within a defined period of a customer, auditor or regulator request.

13.3 **SHALL** AI governance be reviewed at least annually by a function independent of the team that builds and operates the AI systems, and the findings SHALL be reported to the accountable governing body.

13.4 **SHOULD** The organization obtain independent third-party assessment of its AI systems or its AI management system and SHOULD state publicly which standard it is assessed against and the date.

13.5 **SHALL** Where the organization has not obtained independent assessment, it states that plainly rather than characterizing internal review as independent and SHALL record a dated plan.

13.6 **SHALL** Responses given in customer assurance questionnaires be traceable to artifacts in the Evidence Register, and the completed questionnaire SHALL be retained as a record of the representations made.

13.7 **SHALL NOT** The organization claim conformity with, or certification against, a standard it has not been assessed against; nor represent conformity with a management system standard as conformity with a statutory regime.

13.8 **SHOULD** The organization record, at board level, its position on AI insurance, specifically whether it holds affirmative cover for AI output errors, and whether its existing liability, professional indemnity and directors and officers policies contain AI exclusions. Where no affirmative cover exists, the organization SHOULD record that it is self-insured for that exposure.

EVIDENCE

Evidence artifact	Contents / test of adequacy	Owner
Evidence Register	Artifact, location, owner, last refreshed, review cycle, retention. The master index for every other Article.	Governance
Independent review report	Scope, reviewer independence basis, findings, management response, date.	Internal Audit / Third party
Certification or assessment artifact	Certificate, scope statement, validity dates, or a dated plan where none held.	Governance
Completed questionnaire archive	Every customer assurance response issued, with the evidence relied upon.	Legal / Sales
Insurance position record	Affirmative AI cover status; exclusions identified in current policies; board acknowledgement.	Finance / Legal

Part Three — The conformance ladder

Risk assessment across the industry commonly expresses findings on a 4-tier scale, from Low to Critical Risk. A standard that uses different vocabulary makes its adopters, and the AI agents working for them, translate under pressure, which is when overstatement happens.

This standard therefore defines 4 conformance levels that map directly onto that vocabulary. The point isn't to grade organizations. It's to let an organization, or an AI agent acting for it, determine, per Article, which answer it is entitled to give, and to see the specific artifact that would move it up one level.

Level	Name	Meaning	Equivalent risk rating
L3	Governed	The requirement is enforced by a technical control, evidenced automatically, reviewed on a schedule by someone independent, and the evidence is retained. Failure is detected without anyone looking.	Low Risk
L2	Managed	The requirement is documented, consistently applied, and evidenced — but enforcement depends on process discipline rather than a technical control, or review is internal only.	Medium Risk
L1	Documented	A policy or intention exists. Application is inconsistent, and evidence is produced on demand by reconstruction rather than retained as a matter of course.	High Risk
L0	Ad Hoc	No stated position. The question would be answered from memory.	Critical Risk

Using the ladder honestly

The most valuable discipline this standard can impose is the refusal to claim a level you cannot evidence.
An unsupported Low Risk answer that collapses when the assessor asks for the document costs more credibility than an honest Medium with a dated remediation plan attached. Assessors aren't scoring perfection. They're scoring whether your self-assessment can be trusted. An organization that marks itself down in 2 places and up in 10 is believed. An organization that marks itself Low everywhere is re-examined.

The self-assessment worksheet

Reproduce the following for each of the 13 Articles. Completed, it's a gap analysis, a remediation plan and the working paper behind every assurance answer the organization will give for the next year.

Field	What to record
Article	Number and title.
Current level	L0–L3, with the date assessed.
Basis	The specific artifact or control relied upon. If this field is empty, the level is L0 regardless of what was claimed.

Field	What to record
Gap to next level	The one artifact or control that would move the level up. One single thing.
Owner and target date	Named individual and a date, not a quarter.
Questionnaire answer	The answer this level entitles the organization to give, drafted in advance.
Known qualification	Any nuance that must accompany the answer to keep it accurate. For example, that derived outputs persist in the customer's own environment by design.

Part Four — Deriving company policy

This standard was written to be converted into policy, and this Part is the conversion. It is deliberately opinionated about document structure, because the most common failure in AI policy is a single sprawling document that tries to govern employee behavior, engineering practice, contracting and records management at once, and ends up governing none of them because no single audience reads it.

We recommend 11 instruments. Small organizations may merge them but should preserve the boundary between the employee-facing document and the engineering-facing ones, because they have different readers, different review cycles and different enforcement mechanisms.

Policy instrument	Derived from	Mandatory contents
1. AI Governance and Accountability Policy	Art. 1, 2, 13	Scope and definitions; the governing body and its composition; Accountable Owner model and authority to suspend; approval gates and what may not be approved; escalation tiers; annual review cycle; the Evidence Register and who maintains it; the organization's stated position on independent assurance and insurance.
2. AI Acceptable Use Policy (employee-facing)	Art. 3, 5, 6, 11	What tools are approved and for what; the prohibition on entering privileged, personal or client-confidential material into uncontrolled platforms; the prohibition on personal-account use for organizational work; disclosure duties when AI assists client deliverables; the fact-verification duty before anything leaves the building; consequences. Written for humans, at most 3 pages, reissued when new tools are adopted.
3. AI System Registration Standard	Art. 4	What must be registered and when; the mandatory record fields; the deployment gate; discovery cadence for unregistered use; bill-of-materials generation requirements including autonomy level; treatment of retired systems.
4. AI Impact Assessment Procedure	Art. 3	The functional trigger — never self-classification; the assessment template; differential performance testing method and segments; the mitigate-restrict-withdraw decision rule; human review route; approval and retention.
5. Data Provenance and Model Sourcing Standard	Art. 5	Approved model sources and usage modes; the prohibition on training on customer data absent separate written authorization; acquisition-legality documentation; isolation requirements; the training-versus-processing distinction and the exact language to be used about it; annual provider terms review.
6. AI Records, Retention and Legal Hold Schedule	Art. 6	Records classification of prompts, outputs, logs and tool-call records; per-platform configured retention with verification dates; the AI annex to the legal hold procedure enumerating artifacts and platforms; the deletion-suspension mechanism and its test record; human access rules; redaction at ingestion.
7. Third-Party AI and Sub-Processor Standard	Art. 7	Sub-processor register and fourth-party disclosure; prohibited contract language; technical verification of no-training settings; non-human identity inventory and quarterly review; default-on feature change control; incident notification terms; what assurance artifact is accepted from a supplier and who may waive it.
8. Agent Authorization and Emergency Shutdown Procedure	Art. 8	The autonomy scale and who assigns it; tool permission manifests and re-attestation; the action classes requiring human approval; agent identity issuance; the shutdown procedure with triggers, authority, ordered steps, time to effect and notification; annual shutdown test; code execution isolation requirements.
9. AI Security Standard	Art. 9	Blast-radius containment principle; context isolation architecture; baseline controls; guardrail and egress requirements; adversarial testing scope and cadence; AI incident classes and reporting timelines; vulnerability management for models and connectors; entitlement-scoped retrieval.

Policy instrument	Derived from	Mandatory contents
10. AI Consumption and Cost Control Standard	Art. 10	Quota levels and defaults; circuit breaker thresholds and automatic actions; agent step and recursion bounds; cost attribution model; alerting and ownership; contractual allocation of consumption risk.
11. AI Disclosure and Content Provenance Standard	Art. 11, 12	Interaction disclosure requirements by channel; provenance marking specification and version; visible labelling rules; the prohibition on stripping provenance; system card template and refresh cycle; accuracy thresholds and measurement schedule; version pinning; fact-verification evidence; feedback channel and change log.

The conversion method

1. **Adopt the Articles as the policy framework's clause structure.** Do not renumber. Traceability from a policy clause back to this standard, and forward to an assurance answer, is worth more than an elegant numbering scheme.
2. **Transplant SHALL clauses verbatim as policy requirements.** They are drafted in the third person for exactly this purpose. Change only the defined terms: "the organization" becomes your entity name, "AI Inventory" becomes whatever your register is called.
3. **Convert SHOULD clauses to SHALL where you can meet them, and to recorded exceptions where you cannot.** A SHOULD that is silently ignored is the mechanism by which policy sets rot.
4. **Attach the evidence table to the policy as its compliance annex.** A policy that does not name the artifact proving compliance will be audited by interview, which is the most expensive and least reliable form of audit.
5. **Assign each policy a named owner and a review cycle.** The Acceptable Use Policy should be reviewed whenever a new tool is approved; the rest annually, or on material regulatory change.
6. **Map each policy clause to the external frameworks your customers ask about.** Annex B provides the starting crosswalk. Doing this once saves it being done badly under deadline for every large customer.

> **ONE THING NOT TO DO**
> Do not write an AI policy that regulates employees and stops there. The evidence is unambiguous that the material failures occur in systems the organization deliberately built and deployed, not in staff pasting text into chatbots. Shadow AI is real and Article 4 addresses it, but an acceptable use policy alone is a governance program aimed at the smallest part of the problem.

Part Five — The label and the pledge

Version 1.0 proposed that #AIDoneRight should mean something specific and demanding rather than serve as a marketing badge. It then defined the label as a pledge, which is a category of statement that cannot be failed. This Part fixes that.

5.1 What the label requires

A product, platform or project may bear the #AIDoneRight label when all of the following are true, and the organization is prepared to evidence each of them to a party that does not work for it.

- **Level 2 or above on all 13 Articles.** No Article at L1 or L0. A single unaddressed Article is enough to disqualify, because the Articles aren't independent. An ungoverned supply chain defeats a well-run security program, and an unregistered system is outside every other control.
- **Level 3 on the 4 load-bearing Articles:** Article 1 (Human accountability), Article 4 (Inventory and Bill of Materials), Article 6 (Custody, retention and records) and Article 8 (Bounded agency). These 4 are the ones where process discipline is demonstrably insufficient and a technical control is required.
- **A current Evidence Register.** Every artifact required by the Articles located, owned, dated and retrievable within the organization's stated response period.
- **An independent review within the preceding 12 months.** Third-party assessment where obtainable; at minimum a review by a function independent of the team that builds and operates the systems, reported to the accountable governing body.
- **A named Accountable Owner willing to be identified publicly** in respect of the labelled system.

5.2 What disqualifies

The following disqualify a system from the label regardless of performance against every other requirement.

- **Training on customer data without separate express authorization.**
- **An agent capable of an irreversible or externally visible action without a human approval gate.**
- **No tested capability to stop the system.**
- **Claiming a certification the organization does not hold or presenting a management-system certificate as statutory compliance.**
- **Presenting an AI system as a human being or permitting the organization to characterize it as an actor separate from the organization.**
- **Knowledge of a differential performance disparity affecting a protected population, without mitigation, restriction or withdrawal.**
- **Inability to state the configured retention period for the system's prompts and outputs.**

5.3 The pledge, restated

For a product, platform or project to carry the #AIDoneRight label, its stewards publicly affirm:

- We know exactly what this AI is for, what it must never be used for, and who is accountable for it by name.

- We know where its models came from, what they were trained on, and on what legal basis, and our customers' data is not in them.
- We know every system it can reach and every action it can take, because we enumerated them, and it cannot take a consequential action without a human.
- We can stop it, we have tested that we can stop it, and we know how long that takes.
- We know what it retains, for how long, who can read it, and how we produce it when it is demanded.
- We measure whether it is still correct, we record what we measure, and we have agreed in advance what result would make us switch it off.
- We tell people they are talking to it, we mark what it generates, and we do not hide behind a disclaimer.
- We have written down what we cannot yet do, with a date.
- And we have shown all of this to someone who does not work for us.

AI Done Right asks for responsibility, not perfection. It is the promise that behind every system bearing this label there are humans who care enough to stand in front of it, and a filing cabinet that proves they did more than care.

5.4 On the honest interim

Most organizations reading this will not qualify today, including organizations doing careful work. The right response is to publish the gap, not to soften the standard or quietly self-certify.

An organization that states "we conform to 11 of 13 Articles at Level 2 or above; Articles 13 and 9 are at Level 1, with independent assessment scheduled for the second quarter and an adversarial testing program starting next month" is making a stronger claim than one asserting blanket conformity, because the first can be checked and the second invites the checking to be done adversarially. In a market where an assessor's real question is whether your self-assessment can be trusted, the interim disclosure is not a weakness. It is the product.

Annex A — Enterprise verification crosswalk

This Annex maps the assurance questions enterprise customers now routinely ask of AI-bearing suppliers onto the Articles of this standard. The question themes reflect practice that has become common across the industry. Where the standard requires more than customers ask, we note it; where customers ask something the standard should answer more crisply, that is noted too.

Used properly, this table is read right-to-left: you don't answer the questions and then find evidence. You maintain the evidence and the answers follow.

Question theme	Article	What actually determines your tier
AI policy and governance approach	1	Whether the policy is formal and enforceable with mandatory checkpoints, or advisory. The discriminator is whether a system can reach production without passing a gate.
AI inventory and registry	4	Centralized and automated versus a maintained spreadsheet. Automated discovery is the difference between 2 tiers.
Nature and origin of the models	5	Off-the-shelf, retrieval-augmented, fine-tuned, or custom-trained. Answer precisely and note that agentic capability is assessed separately.
Formal risk assessment	3, 9	Recurring and framework-aligned with adversarial testing and bias audits, versus one-time, versus generic SDLC review. A dated plan to reach the top tier materially improves a middle answer.
Risk classification of the system	2, 3	Whether it makes autonomous high-stakes decisions and whether it touches personal data. Classify on function, not on label.
Intended use and operational boundaries	2	Non-critical assistance, professional tool with expert validation, high-impact decision support, or prohibited use. Your Purpose Statement and Non-Goals are the answer.
Training origin; is client data used	5	The 5-question sequence on training. Decide these answers once, in writing, before anyone is asked them under time pressure.
Model documentation and model cards	5, 11	Whether you hold, or can obtain from upstream, model cards and transparency reports. Naming the providers is a precondition of the top tier.
Legal basis for training data	5	Explicit opt-in and zero client data, versus contractual product improvement, versus implied reuse. Acquisition legality is the exposure, not use doctrine.
Personal data in training	5	Answerable at the top tier only if you also state the training-versus-processing distinction explicitly, since personnel accessing production personal data is processing.
Client data isolation from other tenants	5	Dedicated instance with zero cross-pollination versus aggregated. Must be technically describable, not merely contractual.
Drift and update strategy	12	Automated drift detection versus scheduled versus reactive. Note that provider-initiated model change is the more common event.
Ongoing maintenance and support	1, 12	Whether model-performance SLAs exist and whether AI specialists are reachable. Build engagements frequently have no post-handover AI support arrangement — surface that commercially rather than discovering it at handover.
Security framework	9	Zero-trust with AI-specific protections versus enhanced enterprise baseline. Do not claim differential privacy or continuous red teaming you do not perform.
Prompt injection and instruction leakage	9	Independent guardrail layer with egress blocking versus prompt engineering and delimiters versus reliance on base-model filters.
Model Bill of Materials	4	Automated manifest with hashes and lineage versus static documentation versus model cards only. Generate from build, not by hand.
Isolation of AI-generated code execution	8	MicroVM versus container versus shared interpreter. If your system does not execute generated code at runtime at all, say so — it is stronger than any option offered.
Quotas and rate limits	10	Multi-tiered quotas with automated circuit breakers versus standard rate limiting with manual intervention.

Question theme	Article	What actually determines your tier
Model inversion and membership inference	5, 9	Largely inapplicable where no customer-trained model exists; say that rather than implying controls you do not need.
Third-party AI sub-processors	7	Verified enforcement with zero-retention or no-training clauses versus contractual protection without technical verification. The register must be complete before this answer is given.
Emergency shutdown / kill switch	8	A yes commits you to producing the document. Write it first.
Mandatory human approval for high-stakes actions	8	Binary. Clause 8.3 is the requirement.
Training data retention and disposal	5, 6	Where no training occurs on customer data, state the non-applicability alongside the effective posture rather than selecting a retention policy for data that never exists.
Inference input and output retention	6	Zero retention versus limited abuse monitoring versus extended operational retention. Where derived outputs are materialized into the customer's own platform by design, disclose it here — a later discovery of stored AI output otherwise reads as a contradiction.
Logging and monitoring retention	6	No content logging versus temporary redacted logs versus opt-in improvement logging. Verify that redaction at ingestion is actually implemented before claiming it.
Human review of customer content	6	Zero-sight versus break-glass versus routine sampling. Must match the data protection agreement exactly.
Bias and adversarial testing	3, 9	Continuous red teaming with annual independent third-party review versus scheduled internal testing. The top tier requires a report you can hand over.
Transparency to natural persons	11	Multi-layered with in-product notice and self-service explanation versus documented disclosures and a system card. Surfacing provenance and confidence in the interface is the cheapest route upward.
Hosting and data residency	4, 7	Sovereign or dedicated versus regionalized public cloud. Pair the answer with a personnel-location representation; it neutralizes the follow-the-sun concern that the middle tier is penalized for.
External audit	13	Full certification versus independent technical audit within 12 months versus internal audit only. This is the question most suppliers score worst on and the one most worth investing in.
Ownership of AI output	5, 13	Confirm that your contract language expressly reaches prompts and AI-generated outputs, not only deliverables and work product.
Ownership of prompts	5, 6	Frequently overlooked. Prompts are both intellectual property and discoverable records; ownership and retention should be addressed together.

A NOTE FOR WHOEVER COMPLETES THESE

Enterprise questionnaires contain drafting defects with some regularity: duplicated option text, options distinguished only by risk label with no stated criterion, and an absence of "not applicable" for suppliers who train no models. Raising these politely, in writing, alongside your response costs nothing and shows that the response was read rather than pattern-matched. It also protects you: a defect you flagged cannot later be characterized as an answer you got wrong.

Annex B — Framework crosswalk

Indicative mapping from the Articles to the external instruments customers ask about. This is a starting point for a control mapping exercise, not a certified equivalence; where an article requires more than a framework does, the Article governs.

Article	ISO/IEC 42001 and family	NIST AI RMF	EU AI Act	OWASP / ATLAS / CSA
1 Accountability	Leadership; roles and responsibilities; 42001 Cl. 5	GOVERN 1–3	Art. 22, 26 deployer duties	AICM governance domain
2 Purpose and value	Planning; AI objectives; 42005 impact	MAP 1–2	Art. 9 risk management	—
3 Dignity and impact	42005 AI system impact assessment; 22443 societal	MAP 5; MEASURE 2	Art. 9, 10, 27 FRIA	AICM fairness controls
4 Inventory and BOM	Documented information; asset context	MAP 1	Art. 11 technical documentation; Art. 49 registration	CycloneDX / SPDX; CISA-G7 AI SBOM
5 Provenance and rights	5259 data quality; 8183 data lifecycle	MAP 2; MEASURE 2	Art. 10 data governance; Art. 53 GPAI training-data summary	AICM data domain; LLM04 supply chain
6 Custody and records	24970 AI system logging (pending)	MANAGE 4	Art. 12 logging; Art. 19 record-keeping	AICM logging controls
7 Supply chain	Supplier controls; 42001 Cl. 8	GOVERN 6; MAP 4	Art. 25 value-chain responsibilities	LLM04; ASI04 agentic supply chain
8 Bounded agency	42105 human oversight (pending)	MANAGE 1	Art. 14 human oversight	LLM03 excessive agency; ASI01–ASI10; ATLAS agentic platform
9 Security	27001 / 27002 baseline; 42001 Annex controls	MANAGE 2; MEASURE 2	Art. 15 accuracy, robustness, cybersecurity	LLM01, LLM02, LLM08, LLM10; ATLAS; AICM model security
10 Cost containment	Operational planning and control	MANAGE 3	—	LLM06 unbounded consumption
11 Transparency	42001 Cl. 7.4 communication	GOVERN 4; MAP 5	Art. 13, 50; Transparency Code of Practice	C2PA / Content Credentials
12 Evaluation and drift	25058 / 25059 quality; 42119-2 testing	MEASURE 1–4	Art. 15, 17, 72 post-market monitoring	AICM assurance controls
13 Independent assurance	42006 certification bodies; 42007 (pending)	GOVERN 3	Art. 43 conformity assessment	CSA STAR for AI; HITRUST; AIUC-1

USE THIS TABLE DEFENSIVELY

Do not present a crosswalk as evidence of conformity to the mapped framework. It is a navigational aid for buyers who think in their framework rather than yours. Clause 13.7 prohibits representing mapping as certification, and buyers with mature assurance functions test that distinction deliberately.

Annex C — The Evidence Register

The consolidated list of artifacts this standard requires. An organization that builds nothing else from this document but this register, populated with real locations and real dates, will be able to answer most of what it is asked. It will also discover, in the process of populating it, exactly where it stands.

Art.	Artifact	Refresh	Typical owner	Required for label
1	**AI Accountability Register**	Annual + on change	Governance	Yes — L3 required
1	**Escalation and stop procedure**	Annual	Operations	Yes
1	**Deployment approval minutes**	Per deployment	Governance	Yes
2	**Purpose Statement and Non-Goals**	Per system	Product	Yes
2	**Value Hypothesis and measurement record**	Annual	Product / Finance	Yes
2	**Decommissioning Trigger and review decision**	Annual	Governance	Yes
2	**AI capability claim register**	On publication	Marketing / Legal	Yes
3	**AI Impact Assessment**	Pre-deployment + on change	Governance / Legal	Yes where applicable
3	**Differential performance test record**	Per release	Engineering	Yes where applicable
3	**Human review procedure**	Annual	Operations	Yes where applicable
4	**AI Inventory**	Continuous	Governance / IT	Yes — L3 required
4	**Deployment gate control evidence**	Continuous	Engineering	Yes — L3 required
4	**Shadow AI discovery report**	Quarterly	Security	Yes
4	**AI Bill of Materials**	Per build	Engineering	Yes — L3 required
5	**Model provenance register**	Annual + on change	Engineering / Legal	Yes
5	**Training data legal basis file**	Per dataset	Legal	Yes where applicable
5	**Isolation architecture description**	Annual	Engineering	Yes
5	**Provider terms change review log**	Annual	Legal / Procurement	Yes
6	**AI artifact retention schedule with verified windows**	Semi-annual	Records / Legal	Yes — L3 required
6	**Legal hold procedure — AI annex, with test record**	Annual test	Legal	Yes — L3 required
6	**Human access policy and access log**	Continuous	Security / Legal	Yes
6	**Periodic AI usage audit**	Semi-annual	Records / Security	Yes
7	**AI sub-processor register**	Quarterly	Procurement / Legal	Yes
7	**Non-human identity inventory**	Quarterly	Security	Yes
7	**Technical verification of no-training settings**	Annual	Security	Yes
7	**Supplier terms monitoring log**	Continuous	Procurement	Yes
8	**Autonomy Level declaration**	Per system	Product / Governance	Yes — L3 required
8	**Tool permission manifest with re-attestation**	Quarterly	Engineering / Security	Yes — L3 required
8	**Emergency shutdown procedure with annual test record**	Annual test	Operations	Yes — L3 required
8	**Agent identity register**	Quarterly	Security	Yes — L3 required

Art.	Artifact	Refresh	Typical owner	Required for label
8	**Agent action log**	Continuous	Engineering	Yes — L3 required
9	**AI threat model**	Annual + on change	Security	Yes
9	**Adversarial test report**	Per material release	Security	Yes
9	**AI incident response annex with test record**	Annual	Security	Yes
10	**Quota and circuit breaker configuration**	Continuous	Engineering / FinOps	Yes
10	**Cost attribution report**	Monthly	Finance	Yes
11	**AI interaction disclosure evidence**	Per release	Product	Yes
11	**Provenance marking implementation record**	Per release	Engineering	Yes where applicable
11	**System card**	Per material change	Product	Yes
11	**Disclosure consistency review**	Annual	Legal	Yes
12	**Accuracy measurement record**	Per schedule	Engineering / Product	Yes
12	**Evaluation set and results**	Per release	Engineering	Yes
12	**Version pinning and provider change detection**	Continuous	Engineering	Yes
12	**Fact verification evidence for external documents**	Per document	Delivery / Quality	Yes
12	**Feedback and remediation log**	Continuous	Operations	Yes
13	**Evidence Register (this table, populated)**	Continuous	Governance	Yes
13	**Independent review report**	Annual	Internal Audit / Third party	Yes
13	**Certification or dated assurance plan**	Annual	Governance	Yes
13	**Completed questionnaire archive**	Per response	Legal / Sales	Yes
13	**Insurance position record**	Annual	Finance / Legal	Yes

Annex D — The first 90 days

This is for an organization adopting the standard from a standing start. Order matters. Each step makes the next one cheaper, and the highest-consequence unknowns surface first.

Days 1–15 — Find out what is true

- Stand up the AI Inventory as a single register, even if it is initially a spreadsheet. Populate it from cloud spend, identity grants, code repositories and expense claims before asking anyone to self-declare. The reconciliation gap is the most informative output of the whole exercise.
- For every platform found, record the configured retention period for prompts, outputs and logs. Verify it in the console. Do not record what the contract says.
- Enumerate every OAuth grant, API key and service credential held by an AI integration. Name an owner for each.
- List every AI-related claim the organization currently makes in marketing, proposals and filings.

Days 16–45 — Close the 4 gaps that cannot wait

- Write the emergency shutdown procedure and test it. It's a 1-page document and an afternoon's work. It's also the first thing a serious customer asks for and the first thing you will need in an incident.
- Establish the human approval gate for irreversible and externally visible agent actions. If it cannot be built immediately, reduce the autonomy level until it can.
- Add the AI annex to the legal hold procedure, enumerating artifacts and platforms, and test that automatic deletion can be suspended.
- Assign a named Accountable Owner to every system in the inventory and have each confirm in writing.

Days 46–75 — Build the evidence habit

- Define an accuracy measure and threshold for each production system, and take the first measurement. Where no measure is definable, that is the finding.
- Complete the sub-processor register, including the fourth parties you cannot identify. Record those as unknown rather than omitting them.
- Run a first adversarial test against the highest-consequence system, covering indirect prompt injection and tool misuse at minimum.
- Draft the Acceptable Use Policy and the Governance Policy. 2 documents, not 11, at this stage.

Days 76–90 — Make it checkable

- Populate the Evidence Register with real locations and dates. Empty rows are the remediation plan.
- Self-assess all 13 Articles on the Conformance Ladder, recording the basis for each level and the single artifact that would move it up.
- Have someone independent of the build team review the self-assessment and challenge 2 of the levels at random.

— Record the organization’s insurance position (affirmative AI cover, exclusions in current policies) and take it to the board.

WHAT TO EXPECT

Most organizations completing this sequence discover 3 things in a predictable order: more AI systems than the inventory anticipated; retention windows shorter than assumed; and at least one agent integration holding broader credentials than anyone intended. None of these are signs of a badly run organization. They are the normal condition of an estate that grew faster than its governance, which is every estate. The point of the 90 days is that they become known facts with owners rather than unknown facts with exposure.

Annex E — Top "AI Done Right" initiatives globally, ranked by traction

Version 1.0 of this manifesto closed with a survey of the initiatives worldwide that pursue what this standard calls AI Done Right. This Annex carries that survey forward, corrected and updated to the position as of October 31st, 2026. It's here for one reason: a standard that asks others to evidence their claims should evidence its own. Each initiative has been re-checked against its own publications or reputable secondary reporting. Where sources disagree, the disagreement is flagged in the Note on sources rather than smoothed over, and claims from Version 1.0 that could not be confirmed have been removed.

How the ranking works. Traction is judged on 3 things, in this order: binding force (whether anyone can be held to it), breadth of formal adoption (the governments, organizations or individuals that have committed to it), and momentum over 2025–2026 (new instruments, members, signatories or enforcement). The ranking is CodeBoxx's judgment, not a measurement. It has changed since Version 1.0: the EU AI Act moves to first because its obligations now apply; the Global Partnership on AI rises because it merged with the OECD's AI work and grew its membership by roughly 60 percent since Version 1.0; the World Economic Forum entry is renamed because its responsible-AI work now runs through the AI Governance Alliance; and the 2017–2018 researcher and citizen declarations fall because they are static texts.

Rank	Initiative	Type	Position at October 31st, 2026	v1.0 rank
1	**EU AI Act and Trustworthy AI Guidelines**	Binding regulation	Prohibitions, GPAI and transparency duties apply; high-risk deferred to Dec 2027 / Aug 2028	3
2	**OECD AI Principles**	Intergovernmental recommendation	Revised May 2024; 49 adherents; due diligence guidance Feb 2026	1
3	**UNESCO Recommendation on the Ethics of AI**	UN normative instrument	194 Member States; 77 countries supported; 58 readiness assessments completed	2
4	**Global Partnership on AI**	Intergovernmental partnership	Integrated with the OECD since July 2024; 46 countries plus the EU	6
5	**AI for Good**	UN platform	2026 Summit in Geneva drew participants from 169 countries	7
6	**Partnership on AI**	Multi-stakeholder non-profit	10th year; 100+ partners; Responsible AI Action Registry launched	4
7	**WEF AI Governance Alliance**	Multi-stakeholder alliance	Successor vehicle to the Global AI Action Alliance; 350+ members	5
8	**IEEE Ethically Aligned Design and 7000 series**	Standards and certification	10th year; 7000-series standards; CertifAIEd certification	10
9	**Rome Call for AI Ethics**	Faith and industry pledge	New corporate signatories; papal encyclical on AI (May 2026)	11
10	**Asilomar AI Principles**	Researcher pledge	5,720 signatures; no revision since 2017	8
11	**Montréal Declaration for Responsible AI**	Citizen-built declaration	About 2,800 citizen and 300 organizational signatories; no revision	9

READ THIS RANKING CAREFULLY

Traction is not endorsement, and inclusion is not alignment. None of these initiatives has reviewed or endorsed this standard, and conformance to this standard does not imply conformance to any of them. Where a

customer or regulator asks for alignment with a specific instrument, map to the instrument itself using Annex B and evidence it under Article 13.

1. EU AI Act and the Trustworthy AI Guidelines (2019–)

What it is. Regulation (EU) 2024/1689, the first comprehensive horizontal AI law, built on the 2019 Ethics Guidelines for Trustworthy AI and their ALTAI self-assessment list. The Guidelines defined trustworthy AI as lawful, ethical and robust, and set 7 requirements: human agency and oversight; technical robustness and safety; privacy and data governance; transparency; diversity, non-discrimination and fairness; societal and environmental well-being; and accountability. The Act turns that agenda into law: it prohibits a short list of practices, regulates high-risk uses, sets obligations for general-purpose AI models, and imposes transparency duties on AI that interacts with people or generates content.

Origins. The Guidelines were written by the European Commission's High-Level Expert Group on AI. The Act was proposed by the Commission in April 2021, adopted by the European Parliament and Council in 2024, and entered into force on August 1st, 2024. It was amended by the Digital Omnibus on AI, Regulation (EU) 2026/1744, proposed on November 19th, 2025, adopted by Parliament on June 16th, 2026 and by the Council on June 29th, 2026, and in force since July 27th, 2026.

Traction at October 31st, 2026. The prohibitions and the AI literacy duty have applied since February 2nd, 2025, general-purpose AI model obligations since August 2nd, 2025, and the Article 50 transparency obligations since August 2nd, 2026, with generative systems already on the market given until December 2nd, 2026 to add machine-readable marking. The high-risk regime is deferred: stand-alone Annex III systems (employment, credit, education, biometrics and similar) to December 2nd, 2027, and AI embedded in products regulated under Annex I to August 2nd, 2028. Version 1.0 described the Act as forthcoming; it is now law with live obligations, and its high-risk rules did not take effect on August 2nd, 2026 as originally scheduled.

What it means for #AIDoneRight. This is the only initiative in this Annex that can fine anyone. Article 50 is live, so much of Article 11 of this standard is already a legal requirement for EU-facing systems. Treat the high-risk deferral as preparation time, not relief: the Article 3 impact assessment and the Article 12 monitoring evidence are what the deferred regime will ask for.

2. OECD AI Principles (2019, revised 2024)

What it is. The OECD Recommendation on Artificial Intelligence (OECD/LEGAL/0449), the first intergovernmental standard on AI. It sets 5 values-based principles (inclusive growth, sustainable development and well-being; human rights and democratic values, including fairness and privacy; transparency and explainability; robustness, security and safety; and accountability) and 5 recommendations to governments. Its definitions of an AI system and of the AI lifecycle are now used by the EU, the Council of Europe, the United States and the United Nations in their own frameworks.

Origins. Adopted by the OECD Council in May 2019 with 42 countries adhering, and the basis for the G20 AI Principles of June 2019. Revised in May 2024 to account for general-purpose and generative AI.

Traction at October 31st, 2026. 49 adherents as of January 2026: all 38 OECD members, 10 partner economies and the European Union (the public OECD.AI page still lists 47; see the Note on sources). On February 19th, 2026 the OECD published Due Diligence Guidance for Responsible AI, linking the Principles to the OECD Guidelines for Multinational Enterprises, and it now monitors implementation through the OECD.AI Index. Version 1.0's figure of 1,000+ policy initiatives across 70+ countries could not be confirmed against a current primary source and has been removed.

What it means for #AIDoneRight. The shared definitions matter more than the principles: they are why the scope language of this standard travels across jurisdictions. The 2026 due diligence guidance is the closest the OECD has come to an enterprise playbook, and maps naturally onto Article 7 (supply chain) and Article 13 (independent assurance).

3. UNESCO Recommendation on the Ethics of AI (2021)

What it is. The first global normative instrument on the ethics of AI: values and principles anchored in human rights and dignity, plus 11 policy action areas, including data governance, environment, gender, education, culture and health. It is operationalized through 2 tools: the Readiness Assessment Methodology (RAM), which diagnoses a country's legal, institutional and technical preparedness, and the Ethical Impact Assessment (EIA), which teams apply to individual AI systems.

Origins. Developed through a worldwide consultation led by UNESCO and adopted by its General Conference in November 2021 by 193 Member States. It now applies to all 194 Member States.

Traction at October 31st, 2026. UNESCO has partnered with 77 countries on national AI strategies and policies, and 58 have completed the RAM. The 4th Global Forum on the Ethics of AI, held in Riyadh from 14 to September 17th, 2026 and the first in the Gulf region, marked 5 years of the Recommendation and launched RAM 2.0 and a meta-analysis of 55 national reports produced between April 2023 and March 2026. Earlier Forums were held in Prague, Kranj and Bangkok.

What it means for #AIDoneRight. The EIA is the closest public analogue to the impact assessment required by Article 3 of this standard, and the most credible external reference for its differential-impact and environmental questions. The RAM is a national diagnostic; an organization cannot claim it.

4. Global Partnership on AI (2020; integrated with the OECD in 2024)

What it is. An intergovernmental initiative promoting the responsible development and use of AI, grounded in the OECD Recommendation on AI, that connects an international expert community to policymakers through research, evidence and policy tracking.

Origins. Conceived by Canada and France and launched in June 2020 with 15 founding members: Australia, Canada, the European Union, France, Germany, India, Italy, Japan, Korea, Mexico, New Zealand, Singapore, Slovenia, the United Kingdom and the United States. In July 2024, at the New Delhi Summit, GPAI and the OECD's AI work were combined into an integrated partnership under the GPAI brand, operating under OECD procedures; the 2024 Belgrade Declaration guides its work.

Traction at October 31st, 2026. 46 member countries plus the European Union, against the 29 reported in Version 1.0. Meeting on February 20th, 2026 in New Delhi, on the margins of the India AI Impact Summit, the GPAI Council welcomed Malta and Saudi Arabia as new members and committed to expanding its evidence base, including the OECD.AI Policy Navigator and new data on AI compute and investment.

What it means for #AIDoneRight. GPAI's membership now overlaps heavily with the OECD adherents, so it is better read as the research and evidence arm of the OECD Principles than as a separate instrument. Cite it for evidence; do not cite it as a compliance reference.

5. AI for Good (2017–)

What it is. A year-round United Nations platform, organized by the International Telecommunication Union (ITU), that identifies practical applications of AI to advance the Sustainable Development Goals and connects AI innovators with governments and organizations that need solutions. It increasingly hosts standards work, including on AI and multimedia authenticity, alongside its annual Global Summit.

Origins. Launched in Geneva in 2017 by the ITU with the XPRIZE Foundation and UN sister agencies, and co-convened with the Government of Switzerland. It now counts more than 50 UN partners.

Traction as of October 31st, 2026. The 2026 Global Summit, held at Palexpo in Geneva from 7 to 10 July, programmed more than 1,000 speakers and brought together participants from 169 countries. It ran alongside the inaugural UN Global Dialogue on AI Governance and introduced paid registration for the first time, with participation remaining free for ITU Member States, UN agencies, civil society and under-18s. The organizers report an online community of about 37,000 contributors from more than 180 countries. Version 1.0's attendance figure for 2022 has been removed in favor of current, organizer-reported data.

What it means for #AIDoneRight. AI for Good carries convening power and narrative, not obligation: it is a venue in which to show work, not a benchmark against which to claim conformance. Its authenticity-standards track is worth following for Article 11 content-provenance duties.

6. Partnership on AI (2016)

What it is. A non-profit partnership of academic, civil society, industry and media organizations that develops practical guidance on responsible AI. Its best-known outputs include the Responsible Practices for Synthetic Media framework (February 2023), the Guidance for Safe Foundation Model Deployment (October 2023), work on AI, labor and the economy, and the AI Incident Database, which it announced in 2020.

Origins. Announced on September 28th, 2016 by Amazon, Facebook (now Meta), Google and DeepMind, IBM and Microsoft, with Apple joining shortly afterwards. Rebecca Finlay has led it as CEO since October 2021.

Traction at October 31st, 2026. More than 100 partners across 17 countries. In its 10th year, PAI launched a Labor and Economy Steering Committee in March 2026 to support its Shaping Economic Futures in the AI Era initiative, and used its Partner Forum in Geneva on July 6th, 2026, held on the margins of AI for Good and the UN Global Dialogue, to introduce a Responsible AI Action Registry and a Global AI Progress Report. Version 1.0's claims about a 2020 cross-industry COVID-19 endorsement and the composition of PAI's board could not be confirmed and have been removed.

What it means for #AIDoneRight. PAI's synthetic-media and foundation-model guidance are directly usable inputs to Articles 9 and 11. If the Responsible AI Action Registry matures, it becomes a public place to evidence commitments, which is exactly the behavior Article 13 asks for.

7. World Economic Forum's Global AI Action Alliance (2021) and AI Governance Alliance (2023)

What it is. The Forum's flagship responsible-AI initiative, run through its Centre for the Fourth Industrial Revolution. The AI Governance Alliance (AIGA) brings industry, governments, academia and civil society together in 3 working groups: safe systems and technologies, responsible applications and transformation, and resilient governance and regulation.

Origins. The Global AI Action Alliance (GAIA), the vehicle described in Version 1.0, launched in January 2021 with more than 100 organizations and a Steering Committee co-chaired by IBM's Arvind Krishna and Vilas Dhar of the Patrick J. McGovern Foundation. The Forum launched the AI Governance Alliance in June 2023, following its Responsible AI Leadership Summit, and it is now the Forum's primary vehicle for this work.

Traction as of October 31st, 2026. More than 350 members from over 280 organizations, as last reported by the Forum. Version 1.0 anticipated a responsible-AI mark for products emerging from GAIA; no evidence that such a mark was issued could be found, and the claim has been removed.

What it means for #AIDoneRight. The Alliance's strength is executive reach, not standards. Its briefing papers are useful for board-level conversations under Article 1; they are not a conformance reference.

8. IEEE Ethically Aligned Design (2016–2019) and the IEEE 7000 series

What it is. Ethically Aligned Design (First Edition, 2019) sets 8 general principles: human rights, well-being, data agency, effectiveness, transparency, accountability, awareness of misuse, and competence. Its real legacy is the IEEE 7000 series that operationalizes it, including IEEE 7000-2021 (addressing ethical concerns during system design), 7001-2021 (transparency of autonomous systems), 7002-2022 (data privacy process), 7003-2024 (algorithmic bias considerations, released in January 2025), 7014-2024 (emulated empathy) and 7014.1-2026. The IEEE CertifAIEd program, launched in 2021, certifies both products and individual assessors.

Origins. Produced by the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, launched in 2016, which published Version One of Ethically Aligned Design that year and consolidated hundreds of contributors' input into the 2019 First Edition.

Traction as of October 31st, 2026. 2026 marks the Initiative's 10th anniversary. CertifAIEd reported 167 authorized assessors across 28 nations in late 2024 and continues to run assessor training through 2026, and selected 7000-series standards are available at no cost through the IEEE GET Program.

What it means for #AIDoneRight. This is the most directly engineering-usable entry in the Annex. IEEE 7000's value-based requirements process supports Article 2, and IEEE 7003 supports Article 3. CertifAIEd is one credible route to Article 13 evidence, but it complements rather than replaces ISO/IEC 42001 certification.

9. Rome Call for AI Ethics (2020)

What it is. A short pledge setting out 6 principles (transparency, inclusion, accountability, impartiality, reliability, and security and privacy) across 3 impact areas (ethics, education and rights), under the banner of "algorethics": building ethics into AI from the design stage so that it serves every person and humanity as a whole.

Origins. Signed in Rome on February 28th, 2020 by the Pontifical Academy for Life, Microsoft, IBM, the UN Food and Agriculture Organization and the Italian Ministry of Innovation. Jewish and Muslim leaders signed in January 2023, representatives of Eastern religious traditions in Hiroshima in July 2024, and the Archbishop of Canterbury on behalf of the Anglican Communion. The Call is now stewarded by the Vatican's RenAIssance Foundation.

Traction at October 31st, 2026. Corporate signatories now include Cisco, Qualcomm (June 2025) and Salesforce alongside the founding signatories. The Holy See's AI ethics work reached its highest level on May 25th, 2026, when Pope Leo XIV published his first encyclical, Magnifica Humanitas: On Safeguarding the Human Person in the Time of Artificial Intelligence, the first encyclical focused extensively on AI. In July 2026 more than 200 academics, innovators and Nobel laureates signed a declaration in Rome, inspired by the encyclical, on limits for AI and nuclear arms. Version 1.0 placed Cisco's signature and the interfaith expansion together in 2023; Cisco signed later.

What it means for #AIDoneRight. The Rome Call adds no requirement that the OECD or UNESCO instruments lack, but it reaches audiences that secular frameworks do not, and its insistence that AI must

not have the gradual replacement of people as its goal anticipates the workforce concerns in Part One of this standard.

10. Asilomar AI Principles (2017)

What it is. 23 principles covering research issues, ethics and values (including safety, transparency, responsibility and human control) and longer-term issues such as avoiding an AI arms race and managing catastrophic risk.

Origins. Drafted at the Beneficial AI 2017 conference in Asilomar, California, in January 2017, and coordinated and published by the Future of Life Institute (FLI).

Traction as of October 31st, 2026. 5,720 signatures recorded by FLI. The text has not been revised since 2017, and FLI's advocacy has moved on to newer instruments: the Statement on Superintelligence (October 2025), which calls for a prohibition on developing superintelligence until there is broad scientific consensus that it can be done safely and controllably and strong public buy-in, and the Pro-Human AI Declaration (March 2026). The principles fall in this ranking because they are a static, historic text, not because they have lost relevance.

What it means for #AIDoneRight. Cite Asilomar for the provenance of ideas that every later framework absorbed, such as human control and responsibility, rather than as a current reference.

11. Montréal Declaration for Responsible AI (2018)

What it is. 10 principles (well-being, respect for autonomy, protection of privacy and intimacy, solidarity, democratic participation, equity, diversity and inclusion, prudence, responsibility, and sustainable development) supported by 59 recommendations.

Origins. Université de Montréal launched the co-construction process on November 3rd, 2017, running public consultations and citizen deliberations with partners including Mila, IVADO, CIFAR and the Government of Québec, and published the Declaration in December 2018.

Traction at October 31st, 2026. The Declaration's own register lists about 2,800 citizen signatories and some 300 signatory organizations, including the City of Montréal, Mila, IVADO, the Centre Hospitalier de l'Université de Montréal and Québec's financial markets regulator, the Autorité des Marchés Financiers. It continues to fund equity, diversity and inclusion projects, but no revised edition had been published as of October 31st, 2026.

What it means for #AIDoneRight. The Declaration's method has outlasted its text: deliberately involving affected people in deciding what responsible AI means. That is the practice Article 3 of this standard requires when it asks who is affected differently and whether they were consulted.

Why a standard rather than a statement

The temptation, in revising a manifesto, is to make it more eloquent. That would have been the wrong revision.

Version 1.0 argued that if "AI Done Right" were a certification, the label would have to mean something very specific and very demanding. It was right, and it then declined to do the specifying. In the year since, the specifying has been done for us: by courts assigning liability on the basis of control rather than comprehension, by regulators requiring AI communications to be captured as books and records, by insurers withdrawing from exactly the exposures our systems create, by certification bodies making assurance purchasable, and above all by an industry that has matured in how it uses AI, to the point where AI agents now write, test and ship the systems that affect people, and need guardrails as explicit as any we would give a new engineer.

None of those instruments care what we believe. They ask what we can show. That's how the ethical project matures. An ethics that cannot be audited protects no one but the people who profess it.

> **What Version 2.0 asserts is narrow and, we think, defensible:** that the commitments in Version 1.0 were correct, that each of them has an artifact that proves it, that the artifact is usually cheaper to produce than the argument about whether it is necessary, and that an organization which produces them will find it has finally made its principles legible to everyone else, without compromising them.

The systems will keep changing. Agents will be superseded by something with a different name and the same underlying problem: capability arriving faster than the arrangements for answering for it. The Articles are written to survive that, because they ask organizational questions rather than technical ones: who is accountable, what may it do, where did it come from, what do we keep, how do we prove it.

Version 3.0 will be written when the answers to those questions have changed again. On current evidence, that will be sooner than anyone would like.

#AIDoneRight

CodeBoxx Technology Corporation

Version 2.1 — October 31st, 2026

Note on sources and currency

This document states facts about law, standards and reported incidents as of August 13th, 2026. Several are recent enough to require verification before external use, and a standard that does not say so invites the embarrassment it is designed to prevent.

Verified primary and near-primary sources relied upon

- EU: Regulation (EU) 2024/1689 (AI Act); Regulation (EU) 2026/1744 (Digital Omnibus on AI), adopted July 8th, 2026, in force July 27th, 2026, confirmed against the EUR-Lex record; Commission Code of Practice on Transparency of AI-generated Content (final, June 10th, 2026); EN 18286:2026 (published July 12th, 2026; Official Journal citation not confirmed).
- United States: Executive Orders 14179, 14365 and 14409; FTC proposed policy statement published for comment July 7th, 2026 (91 FR 41638); state statutes as cited: Colorado SB 26-189, Texas HB 149, California SB 53, AB 2013, SB 942 as amended by AB 853, AB 316, Illinois HB 3773 and SB 315; SEC examination priorities (November 17th, 2025); FINRA 2026 Annual Regulatory Oversight Report (December 9th, 2025).
- Canada: "AI for All" national strategy (June 4th, 2026); Bill C-36 (first reading June 15th, 2026); Treasury Board Directive on Automated Decision-Making; Quebec Act respecting the protection of personal information in the private sector, s. 12.1.
- Standards: ISO/IEC 42001:2023, 42005:2025, 42006:2025, 23894:2023, 5338:2023, with 42105 and 24970 at final draft; NIST AI 100-1, AI 600-1, AI 100-2 E2025; OWASP LLM Top 10 (2026) and Top 10 for Agentic Applications (2026); MITRE ATLAS v2026.05; CSA AI Controls Matrix v1.1; C2PA 2.4; CycloneDX 1.7; SPDX 3.0; CISA and G7 "SBOM for AI — Minimum Elements" (May 12th, 2026).
- Cases: Moffatt v. Air Canada, 2024 BCCRT 149; Landgericht München I, 26 O 869/26 (May 28th, 2026, first instance); Bartz v. Anthropic (settlement finally approved late July 2026, Martínez-Olguín J.); Getty Images v. Stability AI [2025] EWHC (November 4th, 2025); Conservation Law Foundation v. Shell Oil (D. Conn., May 18th, 2026); Mobley v. Workday (N.D. Cal., ongoing).
- Annex E: OECD.AI and OECD publications (including the OECD.AI Index technical paper, January 2026, and Due Diligence Guidance for Responsible AI, February 19th, 2026); GPAI Co-Chairs' summary statement (New Delhi, February 20th, 2026); UNESCO Global Forum releases (September 2026); ITU AI for Good; Partnership on AI; World Economic Forum; IEEE Standards Association; romecall.org; Future of Life Institute; the Montréal Declaration register. Positions checked as of October 31st, 2026.

Claims marked as contested or requiring care in this document

- The widely-quoted finding that approximately 95 percent of generative AI pilots produce no measurable profit-and-loss impact is not peer-reviewed and has been credibly criticized on methodology. It is presented here as a signal alongside stronger corroborating data, and should not be quoted as a measurement.
- Survey findings on unregistered AI agents derive from vendor-commissioned research. The direction is corroborated by the independent vulnerability record; the specific percentages should be attributed with their sponsorship disclosed.
- The 2025 randomized finding that experienced developers were 19 percent slower with AI assistance should not be cited as a current fact; the same laboratory's 2026 follow-up indicates improved conditions confounded by selection effects. The durable finding is the perception gap between believed and actual productivity.

- Reporting on insurer withdrawal from AI liability is well established, but at least one named carrier subsequently stated it was not specifically seeking to use the exclusions. Attribute carefully.
- The United States federal trade regulator's July 2026 statement on preemption and accuracy suppression is a proposed statement published for comment, not a final instrument. It is described as such throughout this document and no requirement here depends on it.
- EN 18286:2026 is published as a European standard but had not been cited in the Official Journal at the time of writing, which is what confers presumption of conformity. Do not describe it as conferring that presumption today.
- Reported judicial reasoning in the May 2026 German decision is paraphrased rather than quoted, because no verbatim English rendering could be confirmed against a primary or German-language source. The decision is first instance and subject to appeal.
- The precise ordering of positions 4 through 10 in the 2026 OWASP LLM Top 10 was not confirmable against the primary publication at the time of writing. The claims made here (the first 2 positions, the rise of Excessive Agency into the top 3, the 4-place rise of Unbounded Consumption, and the renaming of System Prompt Leakage to Hidden Context Exposure) are corroborated across sources; exact positions should be checked against the official publication before citation.
- Authoritative eDiscovery guidance on the retention and disposition of generative AI artifacts does not yet exist. The leading standards body began drafting in July 2026 with publication targeted for April 2027. Article 6 is written to map onto that work when it lands but is currently built on case law and regulatory expectation rather than on settled industry practice.
- Sources disagree on the number of adherents to the OECD AI Principles. The OECD's January 2026 OECD.AI Index paper (38 members, 10 partner economies and the European Union) and April 2026 legal-tracker reporting give 49; the public OECD.AI principles page still states 47. Annex E uses 49; recheck against the OECD legal instruments register before external citation.
- The ranking in Annex E is CodeBoxx's judgment of traction, not a published metric. Event attendance, community size and signatory counts are as reported by the organizers or by secondary sources and have not been independently audited. Signatory registers for the Asilomar Principles and the Montréal Declaration are cumulative and include unverified individual entries.

Review cycle.
This document is reviewed every 6 months while the regulatory position remains unsettled, and immediately upon: publication of the EU high-risk harmonized standards; any United States federal preemption statute entering force; publication of the eDiscovery commentary on generative AI information; or a revision of the NIST AI Risk Management Framework.