

Source dossier

The evidence never arrived — 18 confirmed cases where a safety monitor was running and its finding never reached the block, and the duty that asks you to test it is not in the AI Act

Edition 2026-10-20 · narrative: The control is not the compliance artefact; the process that proves it still works is. · compiled 2026-10-20

Claims index

Claim	Statement	Status	Sources
CLAIM-2026-10-20-01	A paper submitted to arXiv on 16 October 2026 and announced in the 19 October cs.SE listing (arXiv:2610.23362) confirms 18 evidence-path failures across 11 LLM or agent safety project groups and four implementation ecosystems, in which a framework continues running while safety-relevant evidence never reaches the point at which a decision is made.	verified	arxiv2026monitorva
CLAIM-2026-10-20-02	Of the 18 confirmed failures, 14 reach enforcement points - seven native action outcomes, six configured enforcement boundaries and one synthetic-consumer consequence - while the remaining four affect verifier, safety-prompt or observability channels.	verified	arxiv2026monitorva
CLAIM-2026-10-20-03	The 18 cases comprise nine established-boundary repairs, eight new-surface policy-coverage gaps and one boundary-only evidence case, grouped by the authors into carrier, adapter, authority and surface families.	verified	arxiv2026monitorva
CLAIM-2026-10-20-04	The paper's method is the canary contract: a parent/fixed confirmation unit that places a known payload only on the changed evidence surface and records monitor visibility together with the downstream decision for action-boundary cases.	verified	arxiv2026monitorva
CLAIM-2026-10-20-05	The cases were reconstructed from public framework repairs, so the sample is biased towards defects already noticed and fixed; 18 is a confirmed count from a single author, not a population rate, and no author affiliation is stated in the work's own front matter.	verified	arxiv2026monitorva

Claim	Statement	Status	Sources
CLAIM-2026-10-20-06	GDPR Article 32(1)(d) requires 'a process for regularly testing, assessing and evaluating the effectiveness of technical and organisational measures for ensuring the security of the processing' - a duty on the testing process rather than on possession of the control, in force since 25 May 2018.	verified	intersoftco2026security
CLAIM-2026-10-20-07	AI Act Article 9 requires a risk management system run as a continuous iterative process throughout the entire lifecycle of a high-risk AI system, with testing performed as appropriate during development and in any event prior to placing on the market or putting into service; Article 9 sits in Chapter III, Section 2.	verified	euaiactexpl2026risk
CLAIM-2026-10-20-08	Regulation (EU) 2026/1744, published 24 July 2026 and in force since 27 July 2026, moved the application of the high-risk obligations for standalone Annex III systems from 2 August 2026 to 2 December 2027 and for Annex I regulated products to 2 August 2028.	verified	eurlex2026regulation
CLAIM-2026-10-20-09	AI Act Article 72 obliges providers of high-risk AI systems to establish and document a post-market monitoring system that actively and systematically collects and analyses performance data across the system's lifetime in order to evaluate continuous compliance with Chapter III, Section 2.	verified	euaiactexpl2026postmarke

Only claims marked **verified** may carry a post. **single-source** may appear as hedged context; **disputed** must not be published at all.

Monitor-Validity Failures: LLM Safety Evidence Paths at Evolving Framework Boundaries

PRIMARY

Publisher	arXiv
Published	2026-10-16
Retrieved	2026-10-20
Cite key	arxiv2026monitorva
URL	https://arxiv.org/abs/2610.23362

Claims drawn from this source. CLAIM-2026-10-20-01, CLAIM-2026-10-20-02, CLAIM-2026-10-20-03, CLAIM-2026-10-20-04, CLAIM-2026-10-20-05

arXiv:2610.23362 — submitted Fri, 16 Oct 2026 12:09:26 UTC (v1, 276 KB). Announced in the Mon, 19 Oct 2026 cs.SE listing. Subject classes: cs.AI, cs.CR, cs.SE.

Author. Junchi Liao (single author). No affiliation appears in the arXiv front matter; the HTML rendering (arxiv.org/html/2610.23362v1) returned HTTP 429 and could not be read, so **no institution is attributed** in the edition.

Subject. Runtime monitors in LLM applications — components that inspect tool calls, traces, streams, detector outputs and verifier inputs. The paper describes a reliability class in which the framework keeps running while safety-relevant evidence never reaches the point at which a decision is made. The motivating example: a regex detector flagged content, and the framework dropped the result before the block decision.

Method — canary contracts. Parent/fixed confirmation units that place a known payload *only* on the changed evidence surface, and record monitor visibility together with the downstream decision for action-boundary cases.

Results.

- 18 evidence-path failures confirmed across 11 LLM or agent safety project groups.
- Four implementation ecosystems.
- Composition: nine established-boundary repairs, eight new-surface policy-coverage gaps, one boundary-only evidence case.
- 14 of the 18 reach enforcement points (E1): seven native action outcomes, six configured enforcement boundaries, one synthetic-consumer consequence.
- The remaining 4 affect verifier, safety-prompt or observability channels.
- Failure families: carrier, adapter, authority, surface.

Argument. Monitor assumptions need regression tests at schemas, streams, tools and adapters.

Limitation recorded by this edition. The cases were reconstructed from public framework repairs, so the sample is biased towards defects somebody had already noticed and fixed. 18 is a confirmed count from one author, not a population rate.

Security of processing - Art. 32 GDPR

SECONDARY

Publisher	Intersoft Consulting
Published	2026
Retrieved	2026-10-20
Cite key	intersoftco2026security
URL	https://gdpr-info.eu/art-32-gdpr/

Claims drawn from this source. CLAIM-2026-10-20-06

Article 32(1) — "Taking into account the state of the art, the costs of implementation and the nature, scope, context and purposes of processing as well as the risk of varying likelihood and severity for the rights and freedoms of natural persons, the controller and the processor shall implement appropriate technical and organisational measures to ensure a level of security appropriate to the risk, including inter alia as appropriate:"

- (a) "the pseudonymisation and encryption of personal data;"
- (b) "the ability to ensure the ongoing confidentiality, integrity, availability and resilience of processing systems and services;"
- (c) "the ability to restore the availability and access to personal data in a timely manner in the event of a physical or technical incident;"
- (d) "a process for regularly testing, assessing and evaluating the effectiveness of technical and organisational measures for ensuring the security of the processing."

Article 32(2) — in assessing the appropriate level of security, account shall be taken in particular of the risks presented by processing, in particular from accidental or unlawful destruction, loss, alteration, unauthorised disclosure of, or access to personal data.

Why it is the binding limb here. 32(1)(d) names a *process* for testing effectiveness, not the possession of a control. A monitor that was installed, configured and running but whose evidence path was severed by a framework change is an untested measure. In force since 25 May 2018; enforced by the data protection supervisory authorities, not by the AI Act market surveillance authority.

Risk Management System - Article 9 - EU AI Act

SECONDARY

Publisher	EU AI Act Explorer
Published	2026
Retrieved	2026-10-20
Cite key	euaiactexpl2026risk
URL	https://artificialintelligenceact.eu/article/9/

Claims drawn from this source. CLAIM-2026-10-20-07

Article 9(1) — "A risk management system shall be established, implemented, documented and maintained in relation to high-risk AI systems."

Article 9(2) — the risk management system is "a continuous iterative process planned and run throughout the entire lifecycle" of the high-risk AI system.

Article 9(6) — testing must confirm that high-risk AI systems "perform consistently for their intended purpose."

Article 9(7) — "Testing procedures may include testing in real-world conditions" (optional).

Article 9(8) — testing "shall be performed, as appropriate, at any time throughout the development process", and "in any event, prior to their being placed on the market or put into service." The development-stage testing is qualified by "as appropriate"; the pre-market testing is not.

Scope note. Article 9 sits in Chapter III, Section 2 — the requirements for high-risk AI systems. Its application therefore follows the high-risk application dates, which Regulation (EU) 2026/1744 moved.

Regulation (EU) 2026/1744 (Digital Omnibus on AI)

PRIMARY

Publisher	EUR-Lex
Published	2026-07-24
Retrieved	2026-10-20
Cite key	eurlex2026regulatio
URL	https://eur-lex.europa.eu/eli/reg/2026/1744/oj

Claims drawn from this source. CLAIM-2026-10-20-08

Regulation (EU) 2026/1744 — the Digital Omnibus amending Regulation (EU) 2024/1689 (the AI Act). Published 24 July 2026; **entered into force 27 July 2026**.

What it moved.

- High-risk, Annex III standalone systems: 2 August 2026 → **2 December 2027**.
- High-risk, Annex I regulated products: 2 August 2027 → **2 August 2028**.
- Machine-readable marking for pre-existing generative AI systems under Article 50(2): 2 August 2026 → **2 December 2026**.
- Member-state regulatory sandboxes: **2 August 2027**.

The high-risk duties that move with those dates include risk management, technical documentation, logging, human oversight, conformity assessment and registration.

What was not moved. Article 4 (AI literacy, since 2 February 2025, wording amended 27 July 2026); Article 5 (prohibited practices, since 2 February 2025); GPAI model obligations (Article 53, since 2 August 2025); Article 50 transparency (since 2 August 2026), with the single 2 December 2026 concession above.

Post-Market Monitoring by Providers - Article 72 - EU AI Act

SECONDARY

Publisher	EU AI Act Explorer
Published	2026
Retrieved	2026-10-20
Cite key	euaiactexpl2026postmarke
URL	https://artificialintelligenceact.eu/article/72/

Claims drawn from this source. CLAIM-2026-10-20-09

Article 72(1) — "Providers shall establish and document a post-market monitoring system in a manner that is proportionate to the nature of the AI technologies and the risks of the high-risk AI system."

Article 72(2) — the system "shall actively and systematically collect, document and analyse relevant data which may be provided by deployers or which may be collected through other sources on the performance of high-risk AI systems throughout their lifetime", so as to "allow the provider to evaluate the continuous compliance of AI systems" with the requirements in Chapter III, Section 2.

Scope note. The duty binds *providers of high-risk AI systems*. The explorer page states Article 72 applies from 2 August 2026 under Article 113, and it shows two conflicting amended versions of paragraph 3 without marking which is current — which is itself a reason not to quote a date from this page. Because the obligation attaches to the high-risk classification, and Regulation (EU) 2026/1744 moved the Annex III high-risk application date to 2 December 2027, the edition does **not** claim a hard application date for Article 72 and cites it only as the second article practitioners reach for.