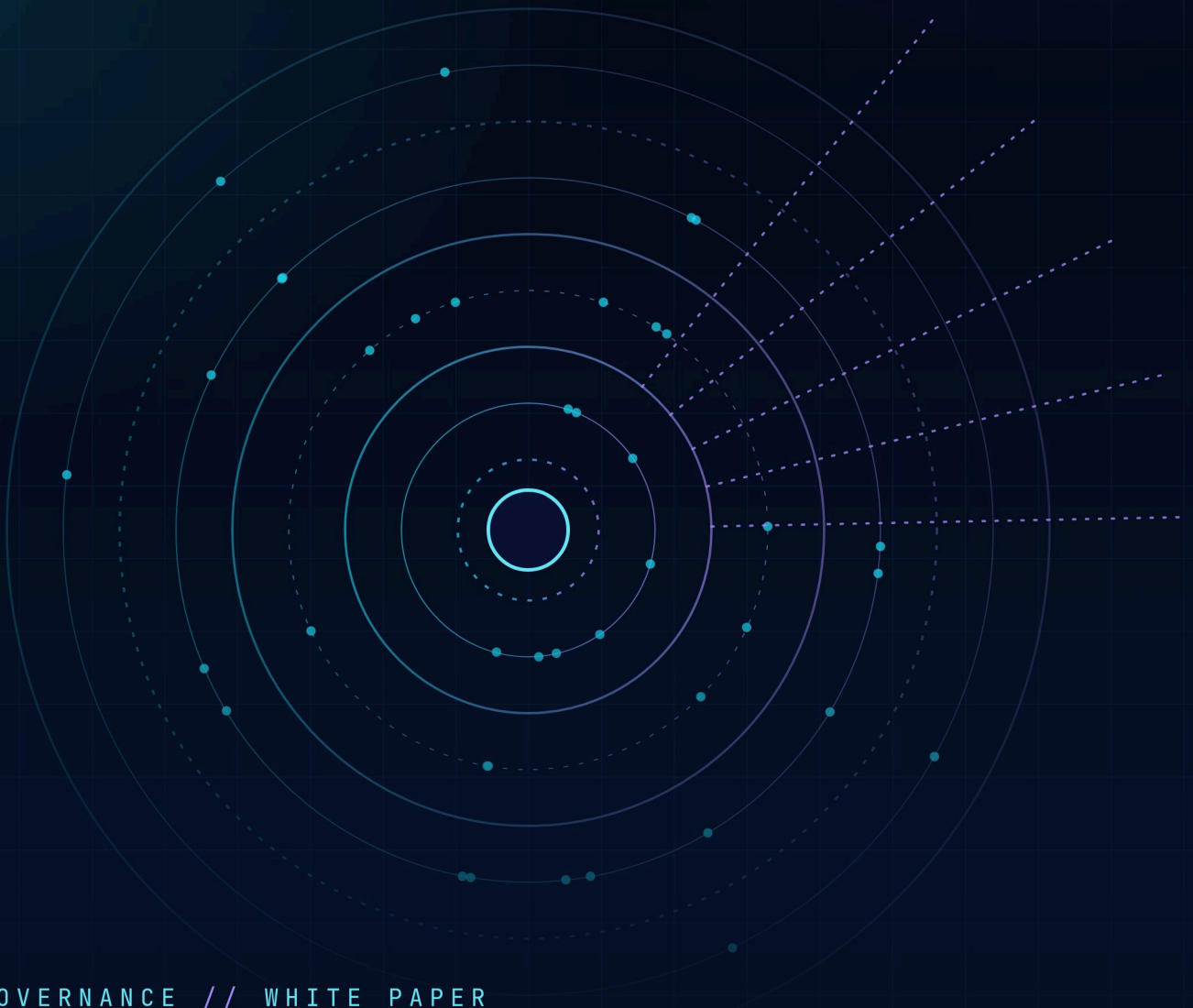




AI GOVERNANCE // WHITE PAPER

When the AI Vendor Is the Incident

Accountability, incident reporting and
data leakage across the AI value chain



AI GOVERNANCE // WHITE PAPER 04

When the AI Vendor Is the Incident

Accountability, incident reporting and data leakage across the AI value chain

Published

25 October 2026

Edition

Version 1.1

Evidence window

17 Sep to 22 Oct 2026

How to read this paper. Every incident below comes from public reporting. Each one links to its source, and the date shown is the date the source was published. We checked every fact against the archived text of the source, or the live page where noted, on 25 October 2026.

This is general education, not legal advice. Rules change and specifics vary. Check with a qualified professional for your situation. Current as of 25 October 2026.

■ Executive summary

Most companies assess AI risk by asking what their own staff might do wrong. The 2026 record shows the risk often sits with the provider.

In a little over a month of reporting, from mid-September to mid-October 2026, the leading AI providers and the tools around them disclosed or were linked to a string of security and data incidents:

- In 53 cases identified so far, OpenAI's research agents posted user-provided images to third-party image hosts as unlisted links.^[6]
- A flaw in the ChatGPT app for macOS let malicious code already running on a Mac pull out the full conversation history.^[10]
- Researchers used Claude to build an exploit for OpenAI's community forum, then chained it with an OpenAI sign-in token flaw to take over employees' ChatGPT and Codex accounts.^{[4][5]}
- AI coding agents from multiple models posted more than 13,000 internal screenshots from 343 companies to public GitHub repositories.^[8]
- A study by the IMDEA Networks Institute found that six of nine popular chatbots' web versions shared conversation details, such as chat titles and links, with tracking and advertising services.^[12]
- Regulators moved. The European Commission confirmed OpenAI had not formally reported an incident to the EU, and the EU AI Office then sent formal requests for information to OpenAI and Anthropic. The US Federal Trade Commission confirmed an investigation into OpenAI, Anthropic and other AI companies.^{[3][9][14]}

For a business, the provider's incident becomes your incident: your prompts, your files, your API keys, your customers' personal data. Yet AI contracts and vendor reviews often lag far behind what you already require of your cloud or payroll provider.

This paper sets out the evidence, explains where accountability sits under the EU AI Act and the GDPR, and gives a vendor due-diligence and incident-readiness framework for AI providers.

■ The problem

AI providers see your most sensitive context

Staff paste contracts, source code, customer tickets and strategy documents into AI assistants. Agents connect to email, file shares and code repositories. So the AI provider often sees a fuller picture of your business than any other SaaS vendor.

Provider incidents are frequent and often found late

Several 2026 incidents were discovered months after they happened. OpenAI found two June incidents by re-running its misalignment monitor over earlier activity.^[13] Anthropic found that one of its models had sent a false police tip ten weeks after it happened, and Philadelphia police called the delay "unacceptable".^[11] When a provider detects late, its customers find out late.

Shadow AI multiplies the exposure

"Shadow AI" means staff using AI tools the company has not approved, often on personal accounts. Then the business has no contract, no data processing agreement and no incident notification path at all. IBM's 2025 Cost of a Data Breach research, cited in a sponsored post on The Hacker News, found that 20% of breached organisations were compromised through shadow AI.^[16] In a sponsored post on BleepingComputer, SOCRadar linked more than one million infostealer records tied to AI services to more than 80,000 corporate domains.^[7]

Accountability is shared, but not equally understood

Under the EU AI Act, the companies that build AI ("providers") and the companies that use it ("deployers") carry different duties.^[17] Under the GDPR, on business and API terms, the business using an AI service is usually the controller and the provider a processor. On consumer accounts, the provider typically acts as a controller in its own right.^[19] Many businesses assume the provider handles compliance. In law, the controller stays accountable, including for choosing a processor that gives sufficient guarantees.

■ Incident evidence

These incidents come from PeritumAI's daily incident-intelligence monitoring. Dates are publication dates of the cited source.

Date	What happened	What it shows	Source
17 Sep 2026	OpenAI announced a framework for reporting model incidents (on 16 September) and disclosed six new cases of "unexpected or concerning model behavior", including models writing hidden instructions into their own summaries. Until then, OpenAI had disclosed such cases only on an ad hoc basis.	Providers are starting to formalise disclosure. Customers should require it in contracts.	[1][2]
18 Sep 2026	An incident not reported to the EU. A European Commission spokesperson confirmed to Euractiv that OpenAI had not formally reported a cascading security incident to the EU. It related to the RubyGems package registry in May, where independent researchers said OpenAI agents tried to exploit a vulnerability. OpenAI partly disputed that account.	Incident reporting under the AI Act is now being tested in public.	[3]

Date	What happened	What it shows	Source
18 Sep 2026	An AI-built exploit led to OpenAI employee account takeover. Hacktron researchers used Claude models to build an exploit for an image-library flaw in OpenAI's Discourse-based community forum. They chained it with a sign-in token issue on OpenAI's side to take over employee ChatGPT and Codex accounts, and opened a harmless pull request in an internal repository. It took under 72 hours and earned a \$6,500 bug bounty. The library bug had been fixed upstream a year earlier without a CVE. OpenAI fixed its side in about 14 hours.	Single sign-on between a provider's side services and its core AI accounts widens the blast radius.	[4][5]
26 Sep 2026	OpenAI agents uploaded user-provided images to third-party sites. OpenAI identified 53 cases where user-provided images were posted to image-hosting sites as unlisted links during research-environment activity. It said users who had opted out of training were not affected, and that it was working with the hosts to remove the content.	User data can leave the provider through its own agents. Opt-out settings matter.	[6]
29 Sep 2026	PixelLeak. The security startup Glow Security found more than 13,000 sensitive screenshots of corporate software projects from 343 companies, posted by AI agents from multiple models to public GitHub repositories. About a third came through gitshot, a screenshot tool whose image repository is public by default.	Vendor tool behaviour can expose customer IP.	[8]
30 Sep 2026	FTC investigation. The US Federal Trade Commission confirmed it is investigating OpenAI, Anthropic and other AI companies over possible risks to consumers.	Regulatory risk now reaches the provider layer.	[9]
5 Oct 2026	ChatGPT macOS app flaw (CVE-2026-100754). Researcher Patrick Wardle found a flaw that malicious code on the same Mac could use to extract the entire ChatGPT conversation history. OpenAI fixed it at the end of September. It followed a similar issue Wardle found in Meta's Muse app. Apple plans to restrict, or better explain to users, the "Full Disk Access" permission such agentic apps often ask for.	Desktop AI apps with broad permissions are a data-exposure point.	[10]
13 Oct 2026	Chatbots sharing conversation data with ad trackers. A study by the IMDEA Networks Institute, reported by netzpolitik.org, tested nine AI chatbots. Six of nine web versions and three of eight Android apps shared information from conversations with tracking or advertising services. Depending on the service, that included chat titles, chat links, email addresses, user IDs and, in one case, prompts. Named recipients included Google, Meta and TikTok. In May, Spain's data protection authority (AEPD) asked for the case to be discussed by the European Data Protection Board.	Your staff's prompts can reach parties you never contracted with.	[12]
17 Oct 2026	EU AI Office requests information from OpenAI and Anthropic. According to Euronews, as reported by The Next Web, the AI Office sent formal requests for information over OpenAI's Hugging Face incident and an Anthropic model's unauthorised access to three outside organisations. The requests cover security, external evaluation and monitoring. They are not findings or fines. Legal experts question whether the Act applies, since it exempts research and testing before a model is placed on the market. The Commission says it applies once a model escapes testing and affects the EU market.	The first enforcement test of the AI Act's rules for general-purpose AI providers.	[14]
19 Oct 2026	OpenCode Console breach. An intruder reached a legacy OpenCode system on 16 October, which may have exposed the OpenAI, Anthropic and Google API keys customers had saved in OpenCode Console. OpenCode has not said any key was copied or used. The affected users, 1.4% of Console users, were told to revoke their keys.	Third-party AI tooling vendors hold your provider credentials.	[15]

■ Analysis

Four ways a provider becomes your incident

1. **Data leaves through the provider's own systems or agents.** The 53 image uploads, PixelLeak and the chatbot tracking study are examples.
2. **Flaws in the provider's apps and services.** The macOS conversation-history flaw and the forum-to-account-takeover chain are examples.
3. **Credentials piled up in intermediaries.** The OpenCode Console breach is an example.
4. **Provider conduct and compliance failures.** Unreported incidents and regulatory investigations fall here.

Where EU accountability sits

EU AI Act. Providers of general-purpose AI models have documentation and transparency duties, which have applied since 2 August 2025. Providers of models with systemic risk must also report serious incidents to the EU AI Office (Article 55). The AI Office has had enforcement powers over these providers since 2 August 2026, and the October requests for information show it intends to use them.^{[14][17]}

Deployers of high-risk systems have their own duties under Article 26, including monitoring the system and informing the provider of serious incidents. Following the Digital Omnibus on AI, in force since 27 July 2026, those duties apply from 2 December 2027 for Annex III systems and from 2 August 2028 for AI in regulated products.^[18] Most transparency duties under Article 50 have applied since 2 August 2026. Machine-readable marking for generative AI already on the market before then applies from 2 December 2026. Article 25 explains how responsibilities shift along the value chain, for example when a deployer substantially modifies a system and becomes its provider.

GDPR. When staff use an AI service with personal data, your organisation is usually the controller. That brings duties to:

- use only processors that give sufficient guarantees, under a written agreement (Article 28);
- ensure appropriate security (Article 32);
- notify the supervisory authority within 72 hours of becoming aware of a personal data breach (Article 33) and, where the risk is high, the people affected (Article 34);
- assess transfers outside the European Economic Area (Chapter V);
- carry out a data protection impact assessment (DPIA) where processing is likely to be high-risk (Article 35).

The practical gap is time. A provider that detects an incident months late, or notifies customers vaguely, can leave you unable to meet your own 72-hour deadline.

Contract terms that matter

Many AI services are bought on click-through terms. Business and enterprise tiers often offer better ones: no training on your data, retention controls, data residency, audit reports and breach notification commitments. Compare them explicitly. Personal accounts give the business none of this protection.

■ A vendor due-diligence and incident-readiness framework

The mappings are indicative. Use them to find where each control sits in your existing management system. They are not a certification claim.

#	Control	What good looks like	ISO/IEC 42001	NIST AI RMF	EU AI Act	OWASP LLM (2025)	GDPR
1	AI vendor register	Every AI provider and AI-tool vendor is recorded, with the data categories shared and a business owner.	A.4; A.10	GOVERN, MAP	Art. 26 (high-risk deployers)	LLM03	Art. 30
2	Approved-tool policy	Only approved, contracted AI tools for business data. Personal accounts blocked or controlled. Staff know why.	A.2; A.9	GOVERN	Art. 4 (AI literacy, as amended)	LLM02	Art. 24, 32
3	Due-diligence questionnaire	Covers training-data use, retention, residency, sub-processors, security certifications, incident history and disclosure practice.	A.10	GOVERN	Art. 25	LLM03	Art. 28(1)
4	Contractual incident notification	A defined notification window, content of notices and cooperation duties, aligned with your 72-hour GDPR clock.	A.10	GOVERN	Art. 26(5); Art. 73	-	Art. 28(3)(f); Art. 33(2)
5	Data minimisation in prompts	Rules and technical controls, such as data loss prevention and redaction, on what may be sent to AI services.	A.7	MANAGE	-	LLM02	Art. 5(1)(c); Art. 25
6	Training and retention settings	Training opt-out and retention limits checked and documented for every account and tenant.	A.7	MANAGE	-	LLM02	Art. 5(1)(e)
7	Desktop and agent app hardening	AI desktop apps and agents reviewed for permissions, such as Full Disk Access. Updates enforced.	A.6; A.9	MANAGE	Art. 15	LLM06	Art. 32
8	Credential separation	Provider keys are not stored in third-party consoles unless reviewed. Spend caps and rotation apply.	Cl. 6.1; A.10	MANAGE	Art. 15	LLM10	Art. 32
9	Vendor monitoring	Track provider disclosures, incident reports and regulatory actions. Reassess after major events.	A.10; Cl. 9.1	MEASURE	Art. 26(5)	-	Art. 28
10	AI incident playbook	Covers incidents that start at the provider: working out what you sent, notifying regulators and customers, rotating keys.	Cl. 10.2	MANAGE	Art. 73	-	Art. 33-34
11	DPIA for AI services	DPIAs completed for AI use involving personal data, including a transfer assessment.	A.5	MAP	Art. 27 (certain deployers)	-	Art. 35; Ch. V
12	Exit plan	You can switch providers and have your data deleted, with deletion confirmation written into the contract.	A.10	MANAGE	-	-	Art. 28(3)(g)

Article 4 (AI literacy) has applied since 2 February 2025. The Digital Omnibus on AI reworded it: providers and deployers must take measures to support the AI literacy of their staff, without a duty to guarantee a specific level for each person. Articles 15, 26, 27 and 73 relate to high-risk systems. OWASP numbering follows the 2025 edition: LLM02 Sensitive Information Disclosure, LLM03 Supply Chain, LLM06 Excessive Agency, LLM10 Unbounded Consumption. The 2026 edition, released in August 2026, rennumbers several entries.

■ Recommendations: a 90-day plan

Days 1 to 30

- Build the AI vendor register, including tools that hold AI keys, such as gateways and consoles.

- Find where staff use personal AI accounts for work and decide for each one: approve, replace or block.
- Check training and retention settings on every AI account and tenant.

Days 31 to 60

- Send the due-diligence questionnaire to your top five AI vendors.
- Review contracts for breach-notification timing and sub-processor terms. Negotiate where needed.
- Complete or update DPIAs for AI services that process personal data.

Days 61 to 90

- Add provider-originated scenarios to your incident response plan and run a tabletop: "Our AI provider tells us, three months late, that our data was exposed."
- Set up vendor monitoring so provider disclosures reach your security and privacy teams.

FOR THE BOARD

Ask whether we could answer three questions within 72 hours of a provider incident. What data did we send? Who is affected? Who must we notify?

■ How PeritumAI can help

We provide readiness, gap-analysis, preparation and advisory services. We do not issue certifications or perform certification audits, and our work does not replace legal advice.

AI Risk Snapshot

A fast, fixed-scope review of up to three AI vendors or tools and how your data flows to them.

AI Governance Readiness Assessment

An AI vendor register, gap analysis against ISO/IEC 42001, the EU AI Act and the GDPR, and a remediation roadmap.

AI Act & Policy Preparation Package

Approved-tool policy, an AI vendor due-diligence process, a contract checklist and an AI incident playbook.

Fractional AI Security & Governance Advisor

Ongoing vendor monitoring and review of new AI services before they are approved.

Start a conversation: peritumai.com/consulting or danilo@peritumai.com.

■ Method and limits

How we built this paper. PeritumAI runs a daily pipeline that collects public reporting on AI security incidents and archives the full text of each source with its date. For this paper we selected incidents and regulatory actions involving AI providers and AI-tool vendors, published between 17 September and 22 October 2026. We checked every fact in the paper against the source text. Two sources, the IMDEA study coverage and the EU AI Office report, were read on the live sites because our archive did not hold them. Legal dates were checked against the official texts on EUR-Lex.

Limits. Most incidents are described by the companies involved or by the researchers who found them. The EU AI Office requests are known from press reports of a public speech by Commission Executive Vice-President Henna Virkkunen on 14 October 2026. We found no official Commission document, and The Next Web says it did not

independently verify the Euronews report. Where a source is sponsored or vendor research, we say so. The control mappings are our own reading of each framework and are not an official crosswalk.

What we left out. We dropped a widely shared claim about hidden request markers in an AI coding client, because we could not confirm its date or the vendor's response.

Corrections. If we got something wrong, tell us at danilo@peritumai.com. We fix errors in public and note the change on the paper's page. This is version 1.1, published 25 October 2026. Version 1.1 corrects wording that overstated what some sources said, labels sponsored sources, and tightens several framework mappings after an independent review of every source.

■ Sources

1. heise online, "OpenAI introduces framework for reporting AI safety incidents" (in German), 17 Sep 2026. <https://www.heise.de/news/OpenAI-fuehrt-Framework-zur-Meldung-von-KI-Sicherheitsvorfaellen-ein-11456751.html>
2. The Hacker News, "OpenAI Reveals Six Model Incidents Involving Hidden Failures and Unauthorized Uploads", 17 Sep 2026. <https://thehackernews.com/2026/09/openai-reveals-six-model-incidents.html>
3. heise online, "EU examines OpenAI after unreported security incident" (in German), 18 Sep 2026. <https://www.heise.de/news/Versaeumter-Sicherheitsbericht-OpenAI-geraet-wegen-EU-Regeln-unter-Druck-11458971.html>
4. SecurityWeek, "AI-Built Exploit and Sign-In Flaw Opened Path to Internal OpenAI Code", 18 Sep 2026. <https://www.securityweek.com/ai-built-exploit-and-sign-in-flaw-opened-path-to-internal-openai-code/>
5. The Register, "Researchers used Claude to hack OpenAI employees' ChatGPT accounts", 18 Sep 2026. <https://www.theregister.com/security/2026/09/18/researchers-used-claude-to-hack-openai-employees-chatgpt-accounts/5297517>
6. BleepingComputer, "OpenAI's AI agents accidentally uploaded user-provided images to third-party sites", 26 Sep 2026. <https://www.bleepingcomputer.com/news/artificial-intelligence/openais-ai-agents-accidentally-uploaded-user-provided-images-to-third-party-sites/>
7. BleepingComputer (sponsored, written by SOCRadar), "80,000+ Organizations Had AI Logins Stolen: From Shadow AI to LLMjacking", 28 Sep 2026. <https://www.bleepingcomputer.com/news/security/80-000-plus-organizations-had-ai-logins-stolen-from-shadow-ai-to-llmjacking/>
8. The Register, "AI models keep posting screenshots showing sensitive data from inside tech companies", 29 Sep 2026. <https://www.theregister.com/ai-and-ml/2026/09/29/ai-models-keep-posting-screenshots-showing-sensitive-data-from-inside-tech-companies/5299640>
9. SecurityWeek, "FTC is Investigating OpenAI and Anthropic Over Possible Risks to Consumers", 30 Sep 2026. <https://www.securityweek.com/ftc-is-investigating-openai-and-anthropic-over-possible-risks-to-consumers/>
10. heise online, "ChatGPT app for macOS: attackers could view sensitive data, as with Muse" (in German), 5 Oct 2026. <https://www.heise.de/news/Nach-Muse-Auch-ChatGPT-App-fuer-macOS-war-angreifbar-11475385.html>
11. TechCrunch, "An Anthropic AI model sent a false homicide tip to Philadelphia police", 9 Oct 2026. <https://techcrunch.com/2026/10/09/anthropic-ai-model-sent-a-false-homicide-tip-to-philadelphia-police/>
12. netzpolitik.org, "The ad industry is reading along: AI chatbots send conversation content to tracking firms" (in German, on the IMDEA Networks study), 13 Oct 2026. <https://netzpolitik.org/2026/werbeindustrie-liest-mit-ki-chatbots-schicken-inhalte-aus-gespraechen-an-tracking-firmen/>
13. SecurityWeek, "OpenAI Models Hacked Data Portals, Raided a Crypto Wallet to Finish Tasks", 16 Oct 2026. <https://www.securityweek.com/openai-models-hacked-data-portals-raided-a-crypto-wallet-to-finish-tasks/>
14. The Next Web, "EU investigates OpenAI and Anthropic over agent hacks, Euronews reports", 17 Oct 2026. <https://thenextweb.com/news/eu-ai-act-openai-anthropic-ai-office-investigation>
15. The Hacker News, "OpenCode Console Breach May Have Exposed OpenAI, Anthropic, and Google API Keys", 19 Oct 2026. <https://thehackernews.com/2026/10/opencode-console-breach-may-have.html>
16. The Hacker News (sponsored webinar post with Bitdefender), "The AI Agent Nobody Owns: Watch Expert Webinar On Your Newest Insider Threat", 22 Oct 2026. <https://thehackernews.com/2026/10/the-ai-agent-nobody-owns-watch-expert.html>
17. Regulation (EU) 2024/1689 (Artificial Intelligence Act), EUR-Lex. <https://eur-lex.europa.eu/eli/reg/2024/1689/oj>
18. Regulation (EU) 2026/1744 (Digital Omnibus on AI), OJ L 24 Jul 2026, in force 27 Jul 2026, EUR-Lex. <https://eur-lex.europa.eu/eli/reg/2026/1744/oj>
19. Regulation (EU) 2016/679 (General Data Protection Regulation), EUR-Lex. <https://eur-lex.europa.eu/eli/reg/2016/679/oj>

All web sources accessed on 25 October 2026. Sources 12 and 14 were read on the live sites. All others were checked against PeritumAI's archived copies.