Deltawall

WHITE PAPER

Evidence-grade AI governance

Demonstrating control over AI usage under GDPR, DORA, NIS2 and the EU AI Act, without surveilling the workforce.

September 2026 · Deltawall B.V. · deltawall.ai

Evidence-grade AI governance

> European regulation increasingly asks organisations to demonstrate control over personal and sensitive data, including when their people use AI. This paper describes what evidence-grade means in practice, maps it to the obligations it supports, and shows how to achieve it without surveilling the workforce.

This paper is general information, not legal advice. Each organisation should assess its own obligations with its counsel and its DPO.

1. What the regulations actually ask for

No European regulation says "govern generative AI prompts" in those words. Several of them, read together, make an unmanaged AI data channel hard to defend.

INSTRUMENT	THE OBLIGATION THAT MATTERS HERE
GDPR, Article 5(2)	Accountability: the controller must be able to **demonstrate** compliance with the data-protection principles, not merely comply.
GDPR, Articles 25 and 32	Data protection by design and by default; appropriate technical and organisational security of processing.
GDPR, Article 17	The right to erasure, which any record about a person must be able to honour.
GDPR, Article 35	A data-protection impact assessment where processing is likely to result in a high risk.
DORA, Regulation (EU) 2022/2554	For financial entities, a documented ICT risk-management framework, applicable since 17 January 2025.
NIS2, Directive (EU) 2022/2555	Cybersecurity risk-management measures for essential and important entities (Article 21).
EU AI Act, Regulation (EU) 2024/1689	AI literacy for staff using AI (Article 4), applicable since 2 February 2025; transparency obligations (Article 50) from 2 August 2026.
Cyber Resilience Act, Regulation (EU) 2024/2847	For manufacturers of products with digital elements, reporting of actively exploited vulnerabilities from 11 September 2026: early warning within 24 hours, notification within 72 hours, final report within 14 days of a fix.

The common thread is **demonstrability**. It is not enough that sensitive data did not leak into an external AI system; the organisation must be able to show, when asked, what controls applied and what they did.

2. Why dashboards are not evidence

Most AI-usage tools answer the demonstrability question with a dashboard: charts of usage, lists of blocked events, screenshots exported to a PDF. These are useful for operations and weak as evidence, for three reasons.

- **They depend on trusting the vendor.** A chart shows what the vendor's database says. Nobody outside the vendor can check that the database was not changed.
- **They depend on trusting the employer.** An administrator with sufficient rights can often edit or delete records. That matters to a regulator, and it matters even more to a works council.
- **They often hold the content.** Logs that store prompts or files to prove what happened create a second copy of exactly the data the control was meant to protect.

Evidence-grade records avoid all three problems.

3. Four properties of evidence-grade records

We use evidence-grade to mean a record with four properties.

1. **Tamper-evident.** Any change to any record, by anyone, is detectable.
2. **Independently verifiable.** A third party can check the record on their own machine, without access to the vendor and without trusting the employer.
3. **Minimised.** The record contains what is needed to prove the decision, and nothing more. In particular, no content.
4. **Erasable without breaking.** A person can exercise their right to erasure without destroying the integrity of everyone else's record.

The first two answer the regulator. The last two answer the data subject and, in Europe, the works council.

4. Hash chains, Merkle seals and signatures, in plain terms

Deltawall achieves tamper-evidence and independent verifiability with three well-established cryptographic techniques.

- **A hash chain.** Each record includes a fingerprint (a cryptographic hash) of the record before it. Changing any record changes its fingerprint, which no longer matches the next record, and so on to the end of the chain. You cannot alter the past without the break showing.
- **Merkle seals.** Records are grouped into batches, and each batch is summarised by a single Merkle root. This makes it efficient to prove that a whole period of records is intact, and to export any bounded period on its own.
- **Signatures.** Each seal is signed with an Ed25519 key. The exported evidence pack carries the public key, so anyone can confirm the seals were produced by the organisation's key

and not forged afterwards.

An exported pack is checked by a **standalone verifier**, a command-line tool that runs offline. It reports whether the chain, the seals and the signature are intact. If a single decision has been altered, for example a Block rewritten as an Allow, verification fails.

5. Minimisation, and Article 17 by salt deletion

A Deltawall record contains a data-class label (for example, IBAN), a salted one-way hash of the match, the destination, the channel, the rule and the outcome. It never contains the prompt, the file, keystrokes or screen content. Detection happens on the device, so the content never reaches Deltawall in the first place.

People are recorded as **pseudonymous tokens**. Linking a token back to a person requires a second approver who is not the requester, and the act of unmasking is itself recorded.

Erasure is handled by **salt deletion**. Each person's hashes are salted with a value unique to them. Deleting that salt makes every record about the person permanently unlinkable, so they become anonymous, while the records themselves, and therefore the hash chain and its seals, remain intact and verifiable. The organisation can honour Article 17 without rewriting history.

6. Works councils: protection without surveillance

In the Netherlands, Germany and France, introducing a system that can monitor behaviour or performance engages the works council:

- In Germany, **BetrVG section 87(1) no. 6** gives the works council co-determination over technical devices designed to monitor employee behaviour or performance.
- In the Netherlands, **WOR article 27(1)** gives the ondernemingsraad a consent right over arrangements for processing employees' personal data and for systems that monitor them.
- In France, the **CSE** must be informed and consulted before means of monitoring employees are introduced.

A control that protects data by watching individuals invites a long negotiation, and often a refusal. Evidence-grade design removes the reasons for it:

- **No individual scoring.** Reporting aggregates to department level. There are no individual leaderboards in the product, and a works-council mode removes individual views entirely.
- **Four-eyes unmasking.** Identifying a person needs two people, and leaves a record.
- **A record nobody can rewrite.** The same tamper-evidence that satisfies the regulator protects the employee: the employer cannot alter the record either.
- **Transparency in the product.** The browser extension's own options page lists the payloads it sent: labels, hashes and metadata, never content.

Deltawall provides a works-council pack written for BetrVG section 87, WOR article 27 and the CSE, with suggested clauses for the works agreement, and a DPIA template written for the organisation as controller.

7. A practical checklist

Use these questions to assess any AI-usage control, including ours.

1. Does detection happen on the device, or is content sent elsewhere to be analysed?
2. Does the record contain content, or only decisions and labels?
3. Can a third party verify the record offline, without access to the vendor?
4. Would changing one historical decision be detectable?
5. Can one person be erased without breaking everyone else's record?
6. Is reporting aggregated, with individual identification gated by a second approver?
7. Does the control require TLS interception or a root certificate on every device?
8. Are the limits of coverage stated in writing, including sites, browsers and file types?
9. Is there a documented, tested path for the CRA's 24-hour, 72-hour and 14-day reporting clock?
0. Can the control's rules be simulated before they are enforced?

Conclusion

Governing AI usage in Europe is a question of evidence as much as control. Records that are tamper-evident, independently verifiable, minimised and erasable let an organisation answer the regulator, the data subject and the works council with the same artefact.

To see an evidence pack exported and verified offline, book a 30-minute demo at **deltawall.ai/demo**, or request early access at **deltawall.ai/early-access**.

Deltawall B.V. · info@deltawall.ai · deltawall.ai