ALDERAI AI · BRIEFING

EU AI Act Article 15: what every AI company must test

Accuracy, robustness & cybersecurity — high-risk obligations enter enforcement 2 August 2026.

Under **Regulation (EU) 2024/1689 (the EU AI Act), Article 15**, providers of high-risk AI systems must design them to achieve an **appropriate level of accuracy, robustness and cybersecurity**, and to perform consistently across their lifecycle. The text explicitly requires resilience against attempts to alter a system's use, outputs or performance by exploiting its vulnerabilities. This is no longer a best-practice — it is a **legal obligation you must evidence in your technical documentation**.

WHAT THE ARTICLE REQUIRES

- Appropriate accuracy, with metrics declared in the instructions for use
- Robustness: technical redundancy, fail-safe design, resilience to errors & faults
- Mitigation of feedback loops in systems that keep learning after deployment
- Cybersecurity resilient to third-party manipulation of use, output or behaviour

ATTACKS IT NAMES DIRECTLY

- **Data poisoning** — manipulation of the training dataset
- **Model poisoning** — tampering with pre-trained components
- **Model evasion** — inputs crafted to make the model err (adversarial examples)
- **Confidentiality attacks** — extraction of data, prompts or the model itself

THE 8-POINT READINESS CHECKLIST

1. **Prompt injection & jailbreaks**
   Direct and indirect injection; system-prompt override resistance.
2. **Data & model poisoning exposure**
   Integrity of training data and third-party / pre-trained components.
3. **Adversarial examples / evasion**
   Perturbed inputs that flip classifications or safety behaviour.
4. **Prompt & system-prompt extraction**
   Confidentiality of instructions, tools and hidden context.
5. **Training-data & secret leakage**
   Membership inference, PII exfiltration, model-inversion risk.
6. **Output robustness & consistency**
   Stable behaviour under paraphrase, noise and edge cases.
7. **Agent & tool-use abuse**
   Excessive agency, RAG poisoning, unsafe function-calling (OWASP LLM Top 10).
8. **Evidence for the technical file**
   Declared accuracy metrics, logs and reproducible robustness tests.

The LLM Red-Team Sprint covers all eight — in two weeks.
Fixed-scope adversarial testing mapped to Article 15 & UAE PDPL, with a compliance gap report and fix roadmap.

ahmedalderai.com
Book an AI Risk Triage call
ahmed@ahmedalderai.com