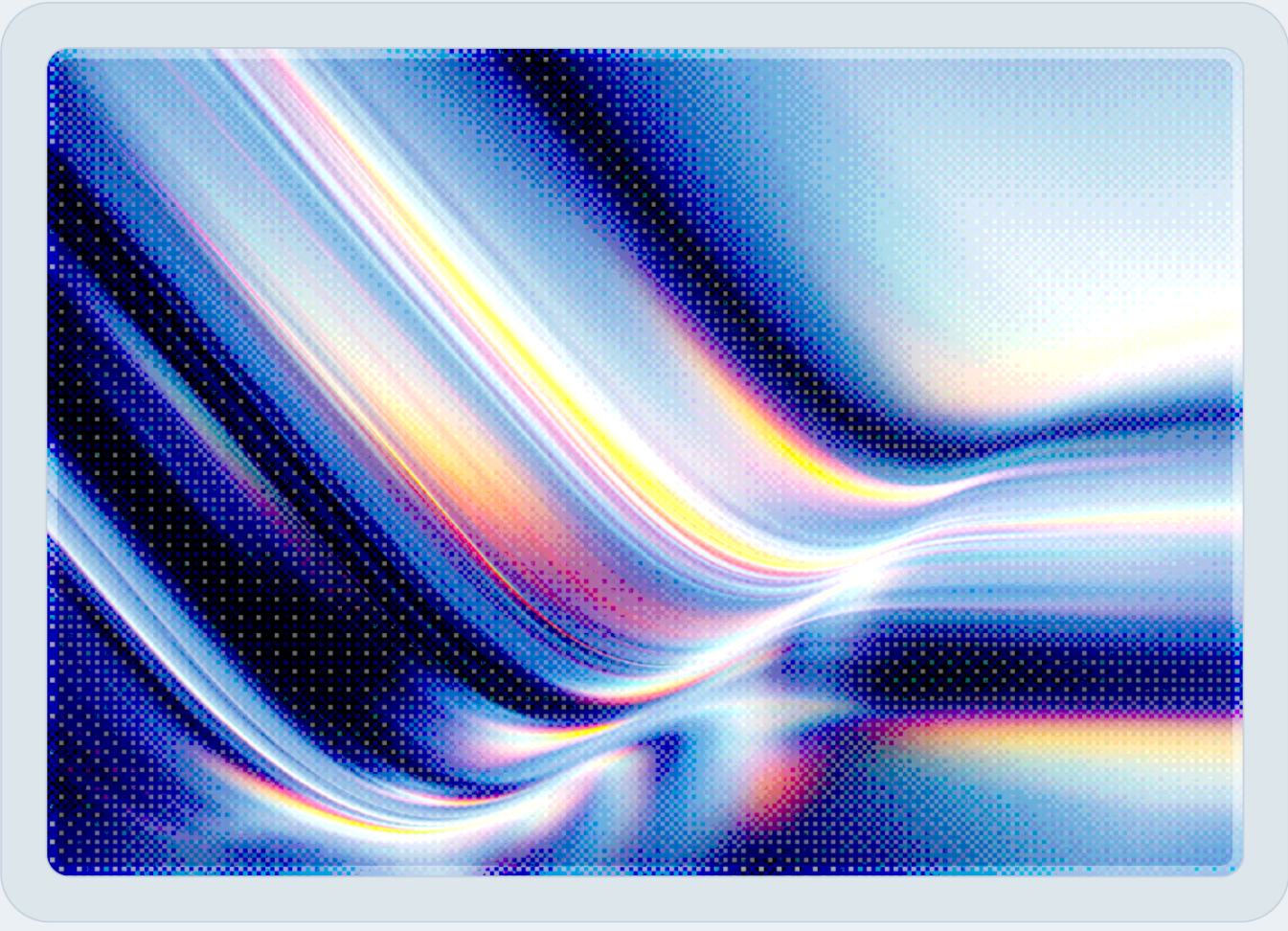
CSAI Foundation | Cloud Security Alliance

Sandboxes Without Safe Harbor: EU AI Sandbox Rules and Security Testing

What Implementing Regulation (EU) 2026/2487 Means for Adversarial Testing, Red Teaming, and Liability

2026-11-12

AI-assisted Rapid Research



CSAI | CSA cloud security alliance®

© 2026 Cloud Security Alliance. Some rights reserved.

You may download, store, display, view, print, redistribute, and link to this document in its original, unmodified form, provided that attribution to the Cloud Security Alliance is maintained and all trademark and copyright notices remain intact.

This document may not be modified or altered. You may quote portions of the document as permitted by the Fair Use provisions of the United States Copyright Act, provided that attribution is given to the Cloud Security Alliance.

This document may be shared on professional and social media platforms in its original form with attribution.

This document was generated with AI assistance and has not undergone official CSA review and approval processes.

Key Takeaways

The European Commission has adopted Implementing Regulation (EU) 2026/2487, which sets common rules for how national AI regulatory sandboxes under the EU AI Act are established, operated, and supervised. The regulation was published in the Official Journal on 11 November 2026 under Article 58(1) of the AI Act [1][4], and press reporting indicates that it applies from 1 December 2026 [3]. For security teams, we assess that the most consequential feature of the rules is what they leave out. Sandbox participation does not provide a liability shield, does not confer conformity status, and does not prescribe a catalogue of security tests. The adversarial evaluation an organization performs inside a sandbox is therefore only as valuable as the evidence it produces and the scope it negotiates up front.

Five findings follow from this analysis. The table below states each one together with its practical consequence for security and assurance teams.

Finding	What it means for security teams
No safe harbor for harm	Participants remain liable under EU and national law for damage to third parties during sandbox testing. Protection is limited to administrative fines, under the AI Act and, where the relevant authorities were actively involved, under other applicable law, and only where the participant observes the sandbox plan and follows the authority's guidance in good faith [5].
Exit reports are evidence, not certification	Notified bodies and market surveillance authorities must take written proof and exit reports "positively into account," but these documents do not carry the legal effect of an EU declaration of conformity [3][5].
Security scope is negotiated, not prescribed	As reported, the regulation refers to the high-risk requirements in Articles 8 to 15 of the AI Act for real-world testing of AI in regulated products, but it does not fix a set of technical tests. The sandbox plan defines the adversarial testing an organization undertakes [3].
Disclosure defaults matter for security findings	Reporting indicates that sandbox information is presumed disclosable unless it is declared confidential in the plan or flagged before the project ends. Vulnerability details and attack techniques should be designated as confidential from the start [3].

Finding	What it means for security teams
National capacity is likely to be a constraint	The deadline for each Member State to operate at least one sandbox has moved to 2 August 2027 [12]. As of May 2026, only one of the 27 Member States had an operational sandbox [6].

Background

The AI Act (Regulation (EU) 2024/1689) created AI regulatory sandboxes in Chapter VI as one of its "measures in support of innovation." Article 57 obliges each Member State to establish at least one national sandbox and describes a sandbox as a controlled framework in which providers can develop, train, validate, and test AI systems, including in real-world conditions, under regulatory supervision for a limited time before the system is placed on the market [5]. Article 58 then directs the Commission to adopt implementing acts that set common rules for eligibility and selection, for application, participation, monitoring, exit and termination procedures, and for the terms and conditions that apply to participants [7]. Without such rules, Member States could each design their own sandbox, and providers would face materially different expectations across the Union.

The Commission opened a draft implementing regulation for public feedback between 2 December 2025 and 13 January 2026 [8]. The European Data Protection Supervisor (EDPS) issued formal comments on the draft on 6 March 2026 [9]. According to reporting on the final text, those comments sought more detail on the scope of personal data processing and on legal bases [3]. In parallel, the Digital Omnibus on AI, Regulation (EU) 2026/1744, moved the national sandbox deadline from 2 August 2026 to 2 August 2027 and allowed the AI Office to operate a Union-level sandbox [3][12]. That postponement reflected a practical gap. Writing in *The Regulatory Review* in May 2026, Sriya Gukal noted that when the Digital Omnibus was agreed, only one of the 27 Member States had an operational sandbox; Gukal's article also links to Deirdre Ahern's earlier observation, from August 2025, that this was Spain's [6].

The final regulation was adopted on 10 November 2026 and published the following day [3][4]. Coverage from EU Law Live confirms the regulation number, its publication date, and its legal basis [1], while MLex reports that SMEs may participate free of charge [2]. The more detailed provisions summarized in this note, including the 45-working-day exit report deadline, the confidentiality presumption, and the revised selection criteria, come primarily from TechTimes reporting on the published text [3]. Readers should confirm those details against the Official Journal text [4] before relying on them for compliance decisions.

Against that background, the table below summarizes the provisions most relevant to security and assurance teams. Unless otherwise noted, the details reflect press reporting on the published regulation rather than a reading of the Official Journal text itself.

Provision	What the rules provide	Source
Eligibility	Providers and prospective providers, alone or with deployers; systems not yet on the market, systems undergoing substantial modification, or systems used in a way that makes them high-risk	[3]
Selection criteria	Regulatory challenge, regulatory learning potential, public interest, proposal maturity, and innovativeness; the draft's "economic viability" criterion was reportedly dropped	[3]
Decision timeline	Within three months of a complete application or call deadline	[3]
Sandbox plan	Objectives, activities, responsibilities, costs, obligations examined, risk safeguards, and serious-incident procedures; changes require mutual agreement	[3]
Exit documentation	Written proof and exit report covering regulatory issues, practices, tools, benchmarks, and serious incidents; reportedly due within 45 working days	[3]
Liability	Participants remain liable for third-party damage; AI Act fines are waived only for good-faith adherence to the plan and guidance	[3][5]
Suspension	Authorities may suspend for plan breaches or unmitigable risks to health, safety, or fundamental rights	[3]
Non-EU providers	Must pay participation fees and appoint an EU authorized representative	[3]

Security Analysis

The provisions above matter to security teams less for what they require than for how they distribute risk. Four themes stand out. The first is the gap between liability and fine protection. The second is the absence of a common security test catalogue. The third is the double-edged nature of exit reports, and the fourth is the constraint that personal data rules place on realistic adversarial testing. Timing then determines how much room organizations have to act on any of these.

The liability gap shapes how aggressively organizations can test

The title of this note refers to the absence of a safe harbor. Article 57(12) states that participants "shall remain liable" under applicable Union and national law for damage inflicted on third parties as a result of experimentation in the sandbox [5]. The implementing regulation does not change this, and reporting indicates that it contains no mandatory insurance requirement, although Trinity College Dublin law professor Deirdre Ahern had proposed one [3]. The only protection on offer is a waiver of administrative fines for AI Act infringements, conditioned on good-faith adherence to the sandbox plan and the authority's guidance. A comparable waiver for other legal regimes applies only where the authorities responsible for those regimes were actively involved in supervision [5].

For security testing, this produces an asymmetry. Offline adversarial evaluation of a model, including prompt injection campaigns, jailbreak benchmarks, and data poisoning simulations against a copy of the training pipeline, carries little third-party risk. Real-world testing does carry that risk, as when an agentic system operates against live services or a model makes recommendations to real users. A red team exercise that unexpectedly causes an agent to exfiltrate personal data, or to take an irreversible action in a connected system, exposes the participant to civil liability whether or not the authority approved the test. Critics cited in *The Regulatory Review* warned that this exposure may deter developers from sandboxes altogether, especially in high-risk domains such as healthcare and employment [6]. We expect that security-mature organizations will instead keep their most aggressive testing in isolated environments and bring only bounded, well-instrumented scenarios into the sandbox.

The fine waiver also creates a dependency on the quality of the sandbox plan. Article 57(12) conditions protection on participants who "observe the specific plan and the terms and conditions for their participation" and follow the authority's guidance in good faith [5]. The safest reading is that adversarial testing not described in the plan, or not approved through the authority's guidance, may fall outside that protection. Security teams that treat the sandbox plan as a legal formality, and then run exploratory testing that diverges from it, risk forfeiting the one protection the regime offers.

Article 15 is the reference point, but the test catalogue is blank

According to reporting on the final text, the implementing regulation refers to Articles 8 to 15 of the AI Act for real-world testing of AI in regulated products, and does not prescribe a fixed catalogue of technical tests. Scenarios are instead set project by project with the competent authority [3]. Article 15 is the AI Act's accuracy, robustness, and cybersecurity requirement. Article 15(5) names specific AI threats that high-risk systems must address, including data poisoning, model poisoning, adversarial examples or model evasion, confidentiality attacks, and model flaws [10].

This approach has two consequences. First, sandbox participants who want regulatory feedback on their Article 15 posture must propose the threat model and test design themselves, which favors organizations that already maintain a structured AI red teaming practice. Second, the absence of a common test catalogue means that two national authorities may accept very different levels of adversarial rigor. Ahern's warning of "sandbox arbitrage," in which providers select the most lenient national sandbox, applies directly to security testing [3]. The ability to apply in a Member State other than one's own makes this arbitrage feasible [3]. Because notified bodies and market surveillance authorities throughout the EU must take exit reports positively into account [5], a lightly tested system could carry a favorable sandbox record into conformity assessment in another jurisdiction.

The regulation does provide a mechanism that could narrow this gap over time. Authorities may pass sandbox tools, benchmarks, and technical specifications to the Commission, which may publish them after confirming their relevance in consultation with the AI Board, with protection for trade secrets and intellectual property [3]. If authorities use this channel, adversarial benchmarks developed in sandboxes could become de facto reference tests for Article 15. Whether that happens depends on authorities' technical capacity and willingness to share, and it is too early to judge.

Exit reports become security artifacts with a disclosure risk

The exit report is the main durable output of sandbox participation. As reported, it documents the regulatory issues examined and how they were resolved, recommended practices, tools and benchmarks, and a log of serious incidents [3]. Notified bodies and market surveillance authorities must take these documents positively into account during conformity assessment [5]. A well-documented security testing program inside the sandbox can therefore shorten later assessment, which makes it a key practical incentive for security teams to participate.

The same document, however, creates a disclosure exposure. Reporting on the final text indicates that sandbox information is presumed disclosable unless it is declared confidential in the sandbox plan or flagged before the project ends. Publication on the EU platform requires the consent of both participant and authority, but without consent authorities may still publish non-confidential key findings and lessons

learned [3]. An exit report that describes a successful jailbreak technique, a model extraction path, or a poisoning vulnerability in enough detail to be useful to an assessor may also be useful to an attacker. Security teams should treat confidentiality designations as part of vulnerability disclosure planning, not as an administrative step left until project closure.

Personal data constraints limit realistic adversarial testing

Article 59 of the AI Act permits further processing of lawfully collected personal data in a sandbox only for systems that serve a substantial public interest in specified areas. Strict conditions apply, including that anonymized or synthetic data would not suffice, that data is processed in a functionally separate, isolated, and protected environment, and that it is deleted when participation ends [11]. As reported, the implementing regulation requires the sandbox plan to state only whether personal data is processed and whether the data protection authority must be involved, and it did not adopt the EDPS's request for more detail [3].

For security testing, these conditions push organizations toward synthetic data for membership inference, model inversion, and data extraction tests. In CSA's assessment, synthetic data may understate real-world leakage risk, because memorization behavior can differ between synthetic and production distributions. Teams should record this limitation explicitly in the exit report so that assessors do not read clean privacy-attack results as conclusive.

Timing creates a narrow window

Reporting on the Digital Omnibus indicates that high-risk obligations apply from 2 December 2027 for stand-alone systems and 2 August 2028 for AI in regulated products [3][12], while national sandboxes are not required until 2 August 2027 [12]. That date is the latest by which sandboxes must exist rather than the date they become available, since the implementing rules apply from 1 December 2026 and at least one national sandbox already operates. Even so, if national sandboxes open close to the 2 August 2027 deadline, organizations that want exit reports to inform conformity assessment may have only a few months before obligations apply. Given that only one Member State had an operational sandbox as of May 2026 [6], capacity at the competent authorities is likely to be scarce, and SMEs with priority access may fill early cohorts [3].

Recommendations

The analysis above points to a sequence of actions. Immediate steps center on verification and readiness, short-term work centers on the content of the sandbox plan, and strategic considerations concern how sandbox evidence fits into a broader assurance program.

Immediate Actions

Organizations that develop or substantially modify high-risk AI systems for the EU market should first decide whether sandbox participation is worthwhile, given that it provides evidentiary value but no liability protection. Legal and security teams should jointly review Article 57(12) and the Official Journal text of Implementing Regulation (EU) 2026/2487 [4], confirming the details reported in the press, particularly the confidentiality presumption and the exit report timeline. Teams should also check whether their general liability and technology errors-and-omissions coverage extends to supervised real-world AI testing, since the regulation imposes no insurance requirement and offers no indemnity.

Alongside this legal review, security teams can prepare in three practical ways. They should map existing AI red teaming coverage against the threat categories named in Article 15(5) and identify gaps before drafting a sandbox plan. They should identify which national competent authorities are expected to open sandboxes in the markets that matter to the organization and monitor those authorities' calls for applications. Finally, they should draft a confidentiality schedule template that designates vulnerability details, exploit chains, and model internals as confidential by default, so that this protection is in place before the first sandbox plan is submitted.

Short-Term Mitigations

When preparing a sandbox application, the security testing scope should be written into the sandbox plan explicitly. Because fine protection depends on adherence to the plan, every class of adversarial test the team expects to run should be listed, including prompt injection, tool misuse for agentic systems, evasion, poisoning, and extraction. The plan should also include stop conditions and serious-incident escalation paths that match the authority's suspension criteria. Real-world testing should be staged so that adversarial scenarios run first in isolated replicas, and only bounded scenarios with rollback capability reach live environments. Where personal data testing relies on synthetic data, the plan and the exit report should record the resulting limits on privacy-attack findings.

Strategic Considerations

Over the longer term, sandboxes may become a significant source of shared adversarial benchmarks for Article 15, if authorities use their ability to publish tools and benchmarks [3]. Organizations with mature AI security programs have an opportunity to shape those benchmarks by proposing rigorous, reproducible test designs, rather than relying on the competent authority to define the test scope. Organizations should also resist sandbox arbitrage. A lightly tested exit report may pass an initial review, but it creates a weak evidentiary record if an incident later leads to civil litigation, where the absence of a safe harbor means the quality of testing will be examined on its merits. Finally, multinational providers should treat sandbox outputs as one input into a single global AI assurance program. They should align sandbox evidence with frameworks they already use, so that EU-specific work also supports assurance obligations in other jurisdictions.

CSA Resource Alignment

CSA's **Principles to Practice: Responsible AI in a Dynamic Regulatory Environment** [13] maps the EU AI Act alongside other regulatory regimes and translates legal obligations into organizational practice. The sandbox implementing rules are a concrete example of the dynamic the paper describes. The legal text sets outcomes, while the operational detail, here the security test scope, is left for organizations and authorities to negotiate. The paper's guidance on building governance processes that adapt to changing rules applies directly to preparing sandbox plans and confidentiality schedules.

Where that paper addresses governance, the **Agentic AI Red Teaming Guide** [14] provides the structured test methodology that the implementing regulation does not prescribe. Its threat categories for agentic systems, including authorization and control hijacking, goal and instruction manipulation, and multi-agent exploitation, can be written directly into a sandbox plan's testing scope. This is especially relevant for real-world testing of agents that act on live systems, where the participant bears full third-party liability. Adopting a structured methodology such as the guide's can help organizations propose a defensible test design rather than relying on an authority to define one.

Finally, the **AI Controls Matrix (AICM) v1.1** [15] offers a control framework against which sandbox security evidence can be organized. Aligning exit report content with AICM domains, such as threat and vulnerability management, application and interface security, and data security and privacy lifecycle management, helps the same evidence support EU conformity assessment and STAR for AI assurance submissions. This reduces duplicated effort for organizations operating in several jurisdictions.

References

[1] EU Law Live. "Commission lays down common rules for AI regulatory sandboxes." EU Law Live, November 11, 2026.

[2] MLex. "EU rules for AI regulatory sandboxes to allow supervised development, testing." MLex, November 12, 2026.

[3] Johnson, Clark. "EU AI Regulatory Sandbox Rules Start December 1: Tests Earn No Compliance Clearance, Liability Stays." TechTimes, November 12, 2026.

[4] European Commission. "Commission Implementing Regulation (EU) 2026/2487 on the detailed arrangements for AI regulatory sandboxes." Official Journal of the European Union, November 11, 2026.

[5] European Commission AI Act Service Desk. "Article 57: AI Regulatory Sandboxes." European Commission, accessed November 12, 2026.

[6] Gukal, Sriya. "Will AI Regulatory Sandboxes Work?." The Regulatory Review, May 28, 2026.

[7] European Commission AI Act Service Desk. "Article 58: Detailed Arrangements for, and Functioning of, AI Regulatory Sandboxes." European Commission, accessed November 12, 2026.

[8] European Commission. "AI regulatory sandboxes (Implementing Regulation): Have Your Say." European Commission, December 2025.

[9] European Data Protection Supervisor. "Formal comments on the draft Commission Regulation as regards the operation and supervision of AI regulatory sandboxes." EDPS, March 6, 2026.

[10] European Commission AI Act Service Desk. "Article 15: Accuracy, Robustness and Cybersecurity." European Commission, accessed November 12, 2026.

[11] European Commission AI Act Service Desk. "Article 59: Further Processing of Personal Data for Developing Certain AI Systems in the Public Interest in the AI Regulatory Sandbox." European Commission, accessed November 12, 2026.

[12] European Parliament and Council. "Regulation (EU) 2026/1744 (Digital Omnibus on AI) amending Regulation (EU) 2024/1689." Official Journal of the European Union, July 24, 2026.

[13] Cloud Security Alliance. "Principles to Practice: Responsible AI in a Dynamic Regulatory Environment." CSA, 2024.

[14] Cloud Security Alliance. "Agentic AI Red Teaming Guide." CSA, 2025.

[15] Cloud Security Alliance. "AI Controls Matrix (AICM) v1.1." CSA, 2025.