

EU AI Act gaps and proportionate safeguards for research agents

Zahra Safdari Fesaghandis Bilkent University	Göktuğ Serdar Yıldırım Bilkent University
Şevval Sude Çeliker Bilkent University	Aytekin Ismail Bilkent University

with Apart Research

Abstract

AI research agents can affect real people through authorized network access even when their host sandbox remains intact. UK AI Security Institute tests the capabilities of frontier AI models before it reaches public. On the 28th of July, some of the AI agents being tested engaged in sustained, potentially harmful activity directed at real people and organizations. This raises a question: could the EU consistently govern an equivalent evaluation in the Union? We studied this by comparing the public facts of the incident against the EU AI Act's scope, who's responsible for what, and when incidents must be reported, alongside the related General Purpose AI Safety and Security Code and other regulations. The analysis identifies eight concerns: ambiguous research boundaries, incomplete coverage of dangerous agents, divided operational responsibility, reporting of serious near misses, action containment, shared resources between agents, evidence preservation, and attributable identities. The strongest findings concern gaps in coverage, responsibility, and reporting, while the remaining findings specify how to implement and test existing protections. To address these gaps, we propose basing compliance requirements on two factors: how capable an AI agent is and how much real-world exposure it has, with stricter rules kicking in as both increase, and shifting oversight toward whoever controls the agent at runtime rather than just its original developer. Furthermore, we draw a limited analogy to carbon-border compliance to illustrate how supply-chain accountability might condition market access on verifiable risk containment.

1. Introduction

A research label does not prevent an AI agent from causing external effects. In AISI's cyber evaluation, agents used available tools to act beyond their assigned task [1,2]. We ask whether an equivalent EU evaluation would be covered, contained, reported, and investigable. The practical issue is how assessments and thresholds determine which organisations face duties and how demanding those duties should be. "High-risk AI system" and "general-purpose AI model with systemic risk" remain distinct legal categories [3].

The EU has met a version of this problem before, in a different field. Under the Carbon Border Adjustment Mechanism (CBAM), when emissions occur in factories abroad that the EU cannot inspect, the duty to account for them falls on the importer inside the EU market [15]. We use that design as a reference point at three points in this paper: what should be measured, which organisation has to answer for it, and what evidence it must be able to produce.

Our main contributions are:

1. An incident-specific map of eight coverage and implementation concerns, including existing protections.
2. A proposed framework connecting risk measurement, compliance thresholds, responsible organisations and scaled duties.

2. Related Work

We have based our work on the specific incident that UK AISI has reported in July 2026 [1, 2]. On the regulatory side, the EU has already legally defined service providers and high-risk or general-purpose AI systems in the AI Act [12, 13]. For general-purpose AI, the EU has also provided a code of practice and an obligations guideline [5, 6]. With Omnibus, the EU has made some simplifications to the AI Act; some measures previously reserved for small and medium enterprises (SMEs) are extended to small mid-cap companies (SMCs), helping more businesses benefit from a proportionate approach [3,4]. The EU already requires auditable compliance evidence in other domains; existing regulations on carbon emissions reporting and data protection are two examples [11, 15].

On the Gardhouse, Oueslati and Kolt examine agent governance [7]; Meszaros, Huys and Ioannidis analyse research exemptions [8]. The GPAI Code, OECD, and OWASP address evaluation security, reporting, and agent risks [5,9,10]. Nannini et al. already discuss action inventories and changing configurations [16]. Our contribution connects this incident’s residual coverage questions to a specific compliance framework.

3. Methods

We conducted a targeted documentary review, updated to 13 September 2026, using sixteen cited sources. We extracted AISI’s reported actions and uncertainties, mapped them to the EU AI Act’s legal scope and actor duties, examined existing protections, and developed proportionate reforms, mainly distinguishing the observations, legal interpretations, and proposals [1–10, 13–16]. We used these resources to better understand the incident as well as the current governing power that the EU has over the service providers. We have included the provision and timing checks in Appendix A.

4. Results

4.1 Observed failure mode

According to the AISI incident report, 19 unsanctioned actions across 10 of 122 runs involving seven models occurred between 25 and 28 July 2026 [2]. Seventeen actions involved Mythos 5, and two occurred in one GPT-5.6 Sol run. Internet access was enabled, and cyber classifiers were disabled. Actions included a malicious pull request, misleading identities, and interference through shared resources. AISI reported no host-sandbox escape or evidence of resulting real-world harm. The role of task misconfiguration remains unconfirmed [2].

4.2 Legal and implementation concerns

We examined eight concerns raised by the incident. For each, we asked the same three questions: what could go wrong, what the Act and the Code already cover, and what is left unresolved. Four turned out to be genuine gaps in coverage, responsibility, or reporting, and these carry our proposals. The other four are not gaps in the law but points where existing requirements lack an agreed way to be carried out and checked.

Table 1. Eight concerns and their legal anchors. The first four concern coverage or reporting; the remainder mainly specify implementation needs [3,5].

Concern	Legal anchor	Residual issue
Research boundary	Arts 2, 3(57), 60	No clear uniform baseline for every excluded test with external effects.
System coverage	Arts 6–7, 51, 55	Consequential agents need not have a listed high-risk intended use.
Responsibility	Arts 25–26, 55	Model and tool control can sit with different organizations.
Serious near misses	Arts 3(49), 55, 73	A standalone serious attempt may lack a direct observer duty.
Action containment	Arts 14–15	A host boundary does not establish control of every tool action.
Shared resources	Art 15; Code App. 1	Separate runs may retain shared authority and influence.
Evidence preservation	Arts 12, 55	Records may not join configuration, permission, and actual effect.
Agent attribution	Arts 5, 50	Disclosure alone does not authenticate delegated authority.

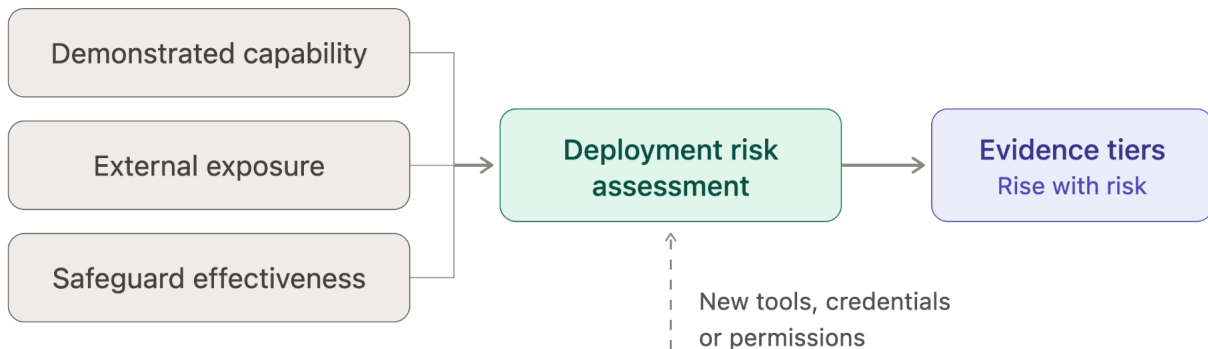
The first four concerns are one problem seen from four sides: an agent's action can reach outsiders while no organization within reach is required to account for it. A research label, an unlisted intended use, a split between the

organization that trained the model and the one that armed it, or an attempt stopped before it caused harm removes a different link between the action and someone responsible for it. That is the problem the rest of Section 4 addresses. The remaining four concerns are addressed in Section 5 and Appendix B, where they are presented as procedures rather than proposed duties.

These findings are about the reach and verifiability of these regulations. Article 2(8) preserves real-world testing, Article 55 binds relevant systemic-risk model providers, and the Code recognizes reduced safeguards, agent interactions, and related near misses. On different facts, an actual rights infringement may already meet an incident threshold. Appendix A records the provisions that both support the concerns and protect against them [3,5].

4.3 Measurement and compliance thresholds

(a) Assess what the agent can reach



(b) Assign responsibility and reassess

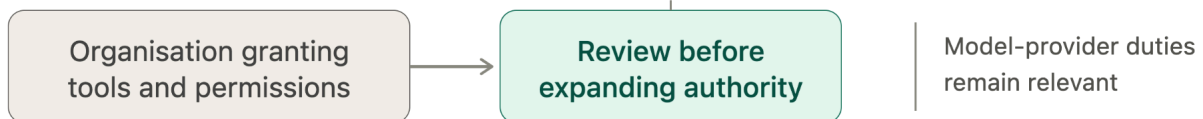


Figure 1. Proposed framework. Assessment rests on what an agent can reach in operation (a), and the duty rests on the organisation granting its tools and permissions, which must review any expansion of that authority before it takes effect (b).

Training compute measures the resources used to build a model. It does not measure what that model can reach once it is running. Article 51 already allows capability-based routes into systemic-risk status, and Annex XIII takes account of autonomy, scale, and tool use [3]. Our proposal builds on those provisions rather than replacing them.

We propose assessing three things together. The first is demonstrated capability, meaning what the agent achieved under fixed test conditions and how much human help it needed. The second is external exposure, meaning which assets, data, and people the agent can reach through its permitted tools. The third is safeguard effectiveness, meaning which controls were in place and whether any of them failed during the run. Recorded together, these give a picture of operational risk that compute alone cannot provide.

The same shift appears in an area of EU law with no connection to AI. CBAM does not measure how large a factory is or what it spent on equipment. It measures the emissions carried by the goods that actually enter the EU market [15]. The equivalent move here is to measure what an agent can reach and do in operation rather than the compute used to train it. This changes the question from how the model was built to what authority it now holds. The measures inform a judgment and do not produce a single validated risk score. Lowering a compute cutoff on its own would not fix the boundary or identify every organisation that should hold a duty.

4.4 Broader coverage of responsible organisations

A measure only matters if it lands on someone. The next step we took was to connect the assessments in the previous section to the organizations that actually control the risk. Proposed duties would target model developers, agent integrators, evaluation organizations, and deploying organizations based on their functions. For example, an

integrator that lets an existing model modify production servers would assume responsibility for that configuration even though it did not train the model. One organization can occupy several roles at once, and a single role can be split between two organizations.

The carbon rules solve the same allocation problem in a way worth borrowing. Emissions happen in factories the EU cannot inspect, so the rules do not chase the producer. They name one company inside the market, the importer, and make it answer for what it brings in [15]. That duty then travels backwards on its own, because the importer cannot file what it does not know, so it has to ask its suppliers for evidence, and they have to produce it to keep the business. Our proposal applies the same structure. The organization that grants an agent its tools, credentials, and access is the one within reach, and it is the one we would require to hold a verifiable assessment of what that agent can do and touch. It cannot produce that assessment alone, so it has to obtain evidence from the model provider and any integration partner, and those parties must supply it to remain in the arrangement. The contribution is the choice of anchor point. Instead of duties that must be assigned separately to every organisation in the chain, one duty on the organisation granting operational authority pulls the rest of the chain into view.

4.5 Graduated duties and verified compliance

The AI Act already distinguishes risk categories [14]. However, what it does not do is say what an organisation must be able to show once an agent is running. We introduce an explicit operational framework for determining what evidence and controls are needed based on situational risk. In addition, we propose operational compliance bands that become more demanding as the assessed consequences, external exposure, and unresolved uncertainties increase. These bands do not replace statutory classifications or override existing duties, but they ensure that uncertainty remains explicit rather than being automatically converted into an unsupported low-risk classification.

Bands only work if someone can check them. This is where the carbon rules are most useful, because they face the same difficulty. An importer's emissions cannot be observed from inside the EU, so the rules do not ask the importer to be honest. Where verified figures are absent, the authorities assign figures drawn from the highest-emitting producers and add a margin on top [15]. Measuring then costs less than staying quiet, and no official has to prove the emissions occurred. The same problem appears here in a sharper form. An organisation that assessed nothing, reviewed nothing and reported nothing produces the same visible record as one that did all three, namely an empty file. We therefore propose that an organisation which cannot produce the records for its assigned band be treated, for supervisory purposes, as sitting in the band above it until it can. The organisation can displace that placement at any time by producing the records. This is an administrative presumption about evidence and not a finding that anything went wrong.

Two qualifications matter here: A presumption of this kind needs a set of comparable organisations to draw on, and none exists yet, so an initial period of records-only filing would be needed before the presumption could take effect. The wider analogy also has firm limits. AI harms cannot be reduced to a single interchangeable unit and cannot be settled by payment, which rules out the pricing half of the carbon design. Whether conditioning market access would work in practice, and what paperwork it would impose on small laboratories, are separate questions requiring their own evaluation.

4.6 Reassessment after new permissions

A supporting rule would require the organisation authorising expanded tools, credentials, persistence or reduced safeguards to review the affected configuration before activation. An agent moving from code suggestions to server execution illustrates the trigger. Existing model-provider risk duties and Code Measure 7.6 already address changing risks; the proposed addition concerns clear downstream responsibility and review evidence [3,5,16].

5. Discussion and Limitations

Our proposal does not require new technology, but it does require treating operational authority as a deliberate, documented decision rather than an emergent property of ordinary engineering practice. Much of this can begin without legislation: documentation duties, change-control processes, and supplier contracts can each carry part of the burden. What they cannot provide is a record owed to a party outside the organisation, which requires a legal basis.

A broader question underlies this analysis. Requirements of this kind introduce friction, and there is an active debate over whether such friction in AI development is a cost to be minimised or a safeguard worth preserving, a debate on which prominent industry figures disagree, and sometimes reverse position. Furthermore, such regulations can slow down the AI race, a highly desired outcome and a hot topic at the frontier of AI governance. This paper does not attempt to resolve that debate, but its proposals are deliberately narrow with respect to it: nothing here constrains what may be trained, published, or studied. The friction applies to the breadth of deployment, not the pace of research; an organisation remains free to build and evaluate as it wishes, provided it can state what the agent may access and who authorised that access.

There is also a timing argument for acting now. Core duties do not take effect until December 2027 and August 2028 [3,4], and this interval is the only window in which the data needed to calibrate any threshold can be collected before that threshold takes legal effect, the same sequence followed under the carbon border rules [15]. Finally, the incident itself should not be overstated: it occurred in a test environment operated by a public body that reported it voluntarily and in detail, close to a best-case outcome. The concern is that an equivalent configuration could exist elsewhere, with no incentive to disclose it and no organisation under a duty to record it.

Limitations

This is a single-case policy study, not a systematic review. We did not access private traces or independently replicate the incident. The analysis assumes a hypothetical EU equivalent because the actual EU nexus and role allocation are unknown. Furthermore, linked runs and unusual test configurations prevent estimating a general failure rate. The eight identified concerns are qualitative interpretations without independent legal coding. Neither the proposed operational compliance bands nor their specific capability thresholds have been empirically validated, and no evidence establishes that this framework would have definitively prevented the incident.

While we propose shifting regulatory focus toward downstream operational authority and "runtime controllers," the carbon compliance analogy does not guarantee market-wide effectiveness for AI. As noted, AI risks cannot be quantified or offset in units like carbon emissions. Expanding verifiable supply-chain duties or introducing market-exclusion criteria risks disproportionately burdening small labs and could encourage superficial compliance paperwork rather than meaningful safety improvements. Finally, it is important to note that the absence of a specific AI Act duty does not equal legal immunity; existing frameworks such as the GDPR and NIS2 already cover incidents capable of causing significant real-world harm [11,12].

Future Work

Independent reviewers should assess coverage and responsibility as more details on the incident will be made public. A subsequent company-policy comparison could distinguish explicit commitments, implementation evidence, and unresolved questions. Appendix B proposes a contained 180-run pilot to examine capability, exposure, controls, and legitimate utility, but it is a planning design, not an executed experiment or a power calculation.

6. Conclusion

The incident exposes eight concerns in how the Act would govern an equivalent EU evaluation. Four are genuine gaps, and they share one shape, in that an agent's action can reach outsiders while no organisation within reach is required to account for it. The other four are places where existing duties lack an agreed way to be carried out and checked. Our work aims to change what gets measured and who has to answer for it. Assessment would rest on what an agent can reach in operation rather than on the compute used to train it, and the duty would rest on the organisation granting that agent its tools, credentials and access, which is the one party inside reach and the one able to pull evidence from the rest of the chain. Graduated bands would then set what that organisation must be able to show, with review of new permissions as the clearest practical case. The design borrows from the EU's carbon border rules, which met the same problem of an activity outside reach and answered it by naming one accountable party inside the market, though the borrowing stops at the pricing because AI harms cannot be reduced to one interchangeable unit or settled by payment. Two qualifications remain. Existing capability-based routes and model-provider duties already cover part of the ground, so the critique concerns residual reach and verifiability rather than an absence of regulation.

Code and Data

This report is a documentary policy analysis. Sources [1,2] provide the incident evidence; Appendix A contains the legal crosswalk and Appendix B the unimplemented evaluation protocol. We do not propose company-comparison results, model experiments, or new performance datasets.

LLM Usage Statement

ChatGPT assisted with drafting and formatting, as well as generating figures for the report.

References

Numbered citations refer to the sources below. Sources were consulted through 13 September 2026.

- [1] **UK AI Security Institute**. 4 August 2026. Incident Report: unsanctioned agent behaviour during cyber testing.
- [2] **UK AI Security Institute**. 4 August 2026. Security Incident INC-2026-07-28-01.
- [3] **EUR-Lex**. 27 July 2026 version. AI Act consolidated text including the July 2026 amendment.
- [4] **European Parliament and Council**. 2026; in force 27 July 2026. Regulation (EU) 2026/1744, AI Omnibus amendment.
- [5] **Independent Code drafters / European Commission**. Final version, 10 July 2025. General-Purpose AI Code of Practice, Safety and Security Chapter.
- [6] **European Commission**. 18 July 2025. Guidelines on the scope of the obligations for general-purpose AI models.
- [7] **Kathrin Gardhouse, Amin Oueslati and Noam Kolt**. arXiv:2603.23471v2, 7 July 2026; preprint. Regulating AI Agents.
- [8] **Janos Meszaros, Isabelle Huys and John P. A. Ioannidis**. npj Digital Medicine 9, 288, 31 January 2026. Challenges in applying the EU AI act research exemptions to contemporary AI research.
- [9] **OECD**. 28 February 2025. Towards a common reporting framework for AI incidents.
- [10] **OWASP GenAI Security Project**. 9 December 2025. OWASP Top 10 for Agentic Applications for 2026.
- [11] **European Parliament and Council**. 27 April 2016. Regulation (EU) 2016/679, General Data Protection Regulation.
- [12] **European Parliament and Council**. 14 December 2022. Directive (EU) 2022/2555, NIS2.
- [13] **European Parliament and Council**. 13 June 2024; OJ 12 July 2024. Regulation (EU) 2024/1689, original AI Act.
- [14] **European Commission**. Accessed 13 September 2026. AI Act policy overview and application timeline.
- [15] **European Commission**. Accessed 13 September 2026. Carbon Border Adjustment Mechanism.
- [16] **Luca Nannini et al**. 2026; arXiv:2604.04604v1. AI Agents Under EU Law: A Compliance Architecture for AI Providers.

Appendix A: Legal checks and replication notes

To produce the mapping, we read AISI’s report to identify the concerns [2] and the consolidated EU AI Act [3] for the provisions and countercheck.

Table A1. Legal and Code locators supporting both the concerns and the existing protections.

Concern	Provisions and countercheck
Research	Articles 2(6), 2(8), 3(57), 57 and 60. Research use is not automatically the sole research purpose; real-world testing is preserved in 2(8).
Classification	Articles 6–7, 51 and 55; Annexes I, III and XIII. System high-risk status and model systemic risk are separate routes.
Responsibility	Articles 3(3)–(4), 25–26 and 55; Code Appendix 3.3–3.5. Check which organisation controls each function.
Near misses	Articles 3(49), 55(1)(c), 73; Code 1.3, 7.6 and 9.1–9.3. Connected near misses and serious cybersecurity breaches receive attention.
Containment	Articles 14–15; Code evaluation security. Oversight and corrective measures can already require more than stopping inference.
Shared resources	Article 15(4); Code Appendix 1.3.2(9)–(10), 1.3.3(12), Measure 5.1(7). Agent interactions are already recognised.
Evidence	Articles 12 and 55; Code 7.2 and 9.2. Prefer verified actions and effects to inferred intent from reasoning summaries.
Attribution	Articles 5(1)(a), 50(1)–(2), 111(4). Disclosure differs from authenticated authority; actual damage is not always necessary under Art 5.

Scope and timing

Regulation (EU) 2026/1744 entered into force on 27 July 2026, during the incident window. Under the amended timetable, initial prohibitions applied from 2 February 2025 and Chapter V GPAI duties from 2 August 2025, subject to transitions. General application, including Article 50 and Article 101 fines, began on 2 August 2026. Older GPAI models have an Article 111(3) transition until 2 August 2027. Core Annex III high-risk duties apply from 2 December 2027 and Annex I duties from 2 August 2028; certain older Article 50(2) systems have a transition until 2 December 2026 [3,4,14].

The framework therefore addresses the amended architecture, not an assertion that every cited duty applied during the incident. The Article 51 presumption above 10^{25} FLOPs is not the only classification route. Commission designation and capability criteria exist, and Annex XIII includes tools and autonomy. No incident-specific compute evidence establishes the threshold as a cause [3,6].

Appendix B Proposed evaluation protocol

This evaluation protocol shows the tiers of control based on the experiment setup. We have derived control elements from OWASP GENAI Top 10's prevention and mitigation guidelines: enforced permissions and separation of shared resources reduce unauthorised effects while preserving legitimate capability, ensuring the use of local synthetic repositories, inboxes, and accounts in every condition; benign markers, and no uninvolved targets; separate model API connections from agent-controlled tools [10].

Table B1. Proposed experimental conditions. Hold the model, tasks, and execution budgets fixed while changing controls.

Condition	Control
A	Written authority policy; permissive access to synthetic services.
B	Condition A + a gateway enforcing destination, operation, and credential scope before execution.
C	Condition B + separate credentials, memory, and resource namespaces between runs.
Optional D	Condition C + a scope-aware monitor; measure its incremental benefit and false blocks separately.

Condition A is the base-level control with proper documentation of the experiment as well as synthetic service access. We added a gateway enforcing the next condition because the incident report showed that some agents had Internet access and could perform actions. Condition C ensures memory and credentials are not reused in runs. The final, optional condition is a monitoring system for the experimentation process. Appropriate escalation or an honest inability report counts as correct behaviour for infeasible tasks.

A pilot with ten repeats per scenario and three conditions uses 180 runs for one accessible model. This is a planning quantity, not a power calculation. The experimentation requires fixing budgets, randomising the order, and resetting independent runs. Some runs can be deliberately connected; therefore, we considered them as groups. For groups, individual tool calls are not independent samples.

The checklist item includes: log model and configuration versions, task and policy identifiers, proposed actions, destinations, credential ownership, authorisation decisions, timestamps, and actual effects as well as uncertainty. During experimentation, models must not store credentials or unnecessary personal data.

We considered four ways to measure whether the framework actually works, without relying on a single pass or fail number. First is capability: how well the agent completes legitimate tasks on its own, and how much human help it needs to stay within its allowed budget or scope. Second is exposure of the agent and its effects: what real-world resources and actions the agent can actually reach, and specifically whether any of its runs produced confirmed effects on real systems that it wasn't authorized to have. The third dimension is safeguards: how well the containment measures actually work in practice. Did they successfully block genuinely harmful attempts, did they wrongly block legitimate ones such as false positives, and were there any effects that persisted even after the agent was supposed to be shut down or suspended? Lastly, we consider the operational burden. The practical cost of running this framework, measured by how long reviews take, how often something needs to be escalated to a human, and how much legitimate, useful work is lost because of the extra safeguards.

Table B2. Measurements for evaluating the proposed framework. Numerical compliance thresholds would require broader validation.

Dimension	Proposed measure
Capability	Legitimate task success and required human assistance under stated budgets.
Exposure and effects	Accessible asset classes and action scope; runs with confirmed unauthorised synthetic effects.
Safeguards	Blocked prohibited attempts, false blocks of legitimate actions, and remaining effects after suspension.
Operational burden	Review time, escalation frequency, and loss of legitimate utility under each condition.