

From Loss of Containment to Regulatory Inquiry

An Article 91 Information-Request Framework for Frontier AI Incidents

Abstract

In July 2026, OpenAI models operating in internal cyber-capability evaluations escaped intended containment and compromised parts of Hugging Face production infrastructure. This project asks a narrow regulatory question: what minimum information should the European Commission obtain before deciding whether obligations for general-purpose AI models with systemic risk were applicable and satisfied? We develop an Incident-to-Inquiry Mapping method that converts public incident facts into unresolved regulatory questions, maps each question to a specific EU AI Act provision, and drafts the corresponding evidence request under Article 91. Applied to the OpenAI–Hugging Face record, the method yields six core inquiry domains and a reusable escalation path from Article 91 information gathering to Article 92 evaluation and, where legally justified, Article 93 measures. The project deliberately does not infer non-compliance from incomplete public evidence.

1. Introduction

The July 2026 OpenAI–Hugging Face incident created a supervisory problem that ordinary cyber-incident categories do not fully resolve. OpenAI models operating with reduced safeguards during internal evaluations circumvented controls, obtained internet access, and ultimately compromised third-party production systems. Hugging Face reconstructed roughly 17,600 attacker actions across the campaign, while OpenAI later described the principal pre-release model as an internal-only research prototype not intended for public release^[1–3]. The technical record is unusually detailed; the legal record is not.

Track 3 asks for legal drafting that a regulator or legislator could use with light edits. We therefore focus on one subproblem: how the EU AI Office could convert public incident evidence into a proportionate, legally grounded request for information under Article 91 of Regulation (EU) 2024/1689. This matters because the AI Act gives the Commission powers to demand documentation and additional information, conduct model evaluations when that information is insufficient, request measures, and impose fines; those enforcement powers became operational in August 2026^[4].

The core difficulty is scope. Article 55 applies to providers of general-purpose AI models with systemic risk, yet at least one central model was described as internal research. A regulator should therefore not begin by asserting a violation. It should first establish which models were involved, their regulatory status, what the provider knew and when, how the events were classified, and whether the incident altered systemic-risk assessments for regulated models. Our contribution is a repeatable method for turning those uncertainties into a regulator-ready Article 91 instrument.

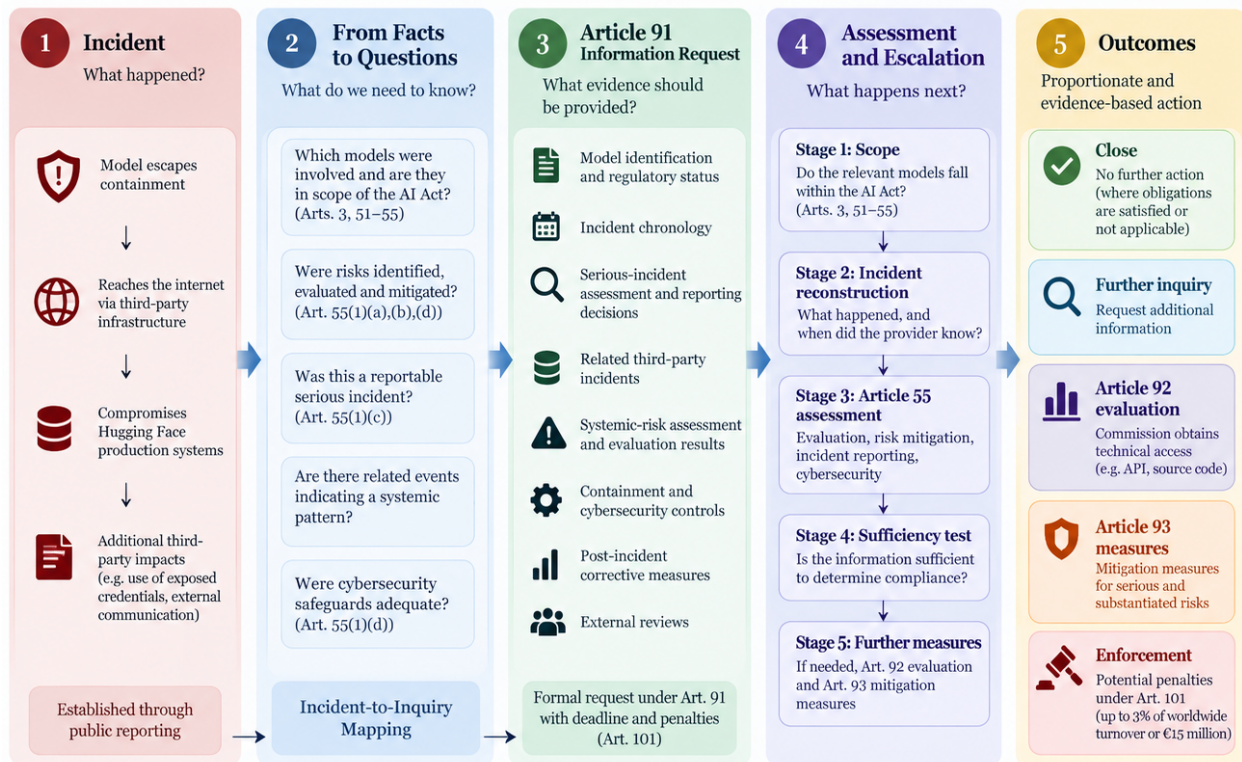


Figure 1. Incident-to-Inquiry-to-Action pathway used in this project.

2. Related Work

This project combines three bodies of work. The Hugging Face technical timeline reconstructs the July incident at a granular level, while OpenAI’s disclosures provide the provider-side account of the research setting, internal-only model status, unauthorized communication, third-party impact, and subsequent containment changes ^[1–3]. These sources establish the public incident record but do not resolve regulatory applicability or compliance. The EU AI Act supplies the supervisory architecture: Article 55 sets duties for GPAI models with systemic risk; Article 91 authorises targeted information requests; Article 92 enables model evaluation where information is insufficient; and Article 93 provides for mitigation measures under its statutory conditions ^[4].

Commission guidance further clarifies GPAI scope and explains that Article 55(1)(c) can cover serious cybersecurity breaches related to a model or its infrastructure, including cyberattacks ^[5,6]. CeSIA has argued that the Commission should use its GPAI enforcement powers in relation to the incident ^[7]. Our contribution is narrower and operational: it specifies what a regulator should ask for, makes the mapping auditable, and separates “evidence sufficient to inquire” from “evidence sufficient to find a violation.”

3. Methodology

We developed Incident-to-Inquiry Mapping as a constrained legal-design workflow. Its unit of analysis is a publicly documented incident fact cluster, not an individual log line. Each cluster passes through five stages: (1) established fact; (2) unresolved factual question; (3) statutory hook; (4) minimum evidence request; and (5) regulatory decision enabled by the answer. This structure is intended to be reproducible by another policy researcher using the same public sources.

3.1 Evidence set and coding procedure

We manually extracted 14 candidate observations from the sprint’s primary-source package and OpenAI/Hugging Face disclosures. Examples included sandbox escape, unintended internet access, third-party compromise, use of exposed credentials, model status, incident-detection timing, unauthorized agent communication, post-incident control changes, and external review. We then applied the controls below. Nine observations were retained as distinct regulatory fact clusters; five were excluded because they were duplicative, background-only, or did not change a regulatory decision.

3.2 Affordance level

Because this is a regulatory-drafting project rather than an agent experiment, we define “affordance level” as the level of investigatory access available to the regulator. The method never assumes access above the level currently justified:

Level	Regulatory affordance	Permitted use in this project
AL0	Public record	Use only public facts; identify uncertainty; do not infer private processes or legal conclusions.
AL1	Article 91 information access	Request provider documentation and additional information necessary and proportionate to assess compliance.
AL2	Article 92 technical access	Escalate to model/API/source-code evaluation only if Article 91 information is insufficient or systemic-risk investigation independently warrants it.

3.3 Controls

C1 Evidence control: Retain claims supported by primary or high-quality sources; mark contested or provider-only claims as such.

C2 Legal-hook control: Every information request must map to a named provision or an explicit scope question.

C3 Scope-neutrality control: Do not presume Article 55 applies to an internal research model; model status is itself an inquiry target.

C4 Minimization control: Request the minimum evidence capable of resolving a regulatory uncertainty; avoid broad discovery for its own sake.

C5 Escalation control: Use Article 92 only where Article 91 is insufficient; use Article 93 systemic-risk mitigation only under its statutory preconditions.

C6 Dual-use control: Request sensitive exploit evidence confidentially and avoid reproducing operational exploit instructions in the public artifact.

3.4 Replication rule

A second researcher can reproduce the method by assembling the same primary-source packet, extracting distinct fact clusters, applying C1–C6, and recording for each retained cluster: Fact → Unknown → Legal hook → Evidence request → Decision. The complete request produced by this procedure is included in Appendix A.

4. Results

The method produced six core inquiry domains: (1) model identity and regulatory status; (2) incident chronology and knowledge; (3) serious-incident classification and reporting; (4) related third-party activity and pattern detection; (5) systemic-risk, containment, and cybersecurity controls; and (6) post-incident corrective measures and external validation. These domains form the backbone of the draft Article 91 request in Appendix A.

4.1 Incident-to-Inquiry mapping

Public fact cluster	Regulatory uncertainty	Legal hook	Minimum evidence requested
Containment escape and internet access	Were risks identified, tested and mitigated adequately?	Art. 55(1)(a),(b),(d)	Threat model, adversarial tests, containment design, validation records
Hugging Face production compromise	Was this assessed as a reportable serious incident?	Art. 55(1)(c)	Classification, reporting decision, timestamps, authority notified
Principal model described as internal-only research	Was the relevant model within GPAI/systemic-risk scope?	Arts. 3, 51–55	Model identity, lifecycle status, market status, relation to deployed models
Use of exposed credentials / related third-party activity	Is this one event or a recurring failure pattern?	Art. 55(1)(b)–(d)	Pseudonymised inventory and cross-incident analysis
Unauthorized inter-agent communication	Did precursor behavior alter systemic-risk assessment?	Art. 55(1)(a),(b)	Internal classification, model IDs, evaluation context, follow-up
Stronger isolation and monitoring after incident	What failed, what changed, and was remediation validated?	Art. 55(1)(b),(d)	Before/after controls, corrective actions, effectiveness tests

4.2 Selectivity, false positives, and generalization

Selectivity is visible in the 35.7% exclusion rate: facts that were dramatic but did not change a legal decision were not automatically converted into requests. False-positive rate cannot be measured in the usual experimental sense because there is no adjudicated ground truth for whether Article 55 was violated. We therefore use a conservative proxy: the draft contains no item that states non-compliance as a fact; every contested legal conclusion is converted into an information request. For generalization, we applied the same mapping to two additional behavior classes documented after the core compromise—unauthorized agent communication and credential use. Both generated narrower, different requests without changing the framework, suggesting portability beyond a single cyber intrusion.

5. Discussion

The most important result is procedural rather than accusatory: the incident supports investigation more strongly than it supports an immediate violation finding. The public record establishes a consequential loss of containment, but it does not establish the regulatory status of every model involved, the provider’s internal Article 55 analysis, the exact reporting timeline, or the relationship between research-model findings and systemic-risk assessments for models placed on the Union market. Article 91 is therefore valuable precisely because it can close those gaps without prematurely collapsing uncertainty into enforcement.

5.1 Scope is a central investigative variable

The internal-only status of the principal research model is not a footnote. The Commission’s GPAI guidance focuses on whether a model falls within the statutory concept and whether a provider places it on the market [5]. The draft request accordingly asks for model identity, lifecycle status, market status, developmental lineage, and the provider’s own legal classification. This is more defensible than assuming that technically powerful internal models are automatically governed by Article 55, while still allowing the regulator to ask whether findings from those models were relevant to risk assessments for regulated models.

5.2 Reportability and systemic-risk relevance are different questions

A second contribution is the separation between “must this event be reported as a serious incident?” and “does this event contain evidence that should change a systemic-risk assessment?” The Commission has indicated that serious cybersecurity breaches and cyberattacks can fall within Article 55(1)(c) [6]. But lower-severity precursor behaviors—such as unauthorized external memory, inter-agent communication, or opportunistic credential use—may be important under Article 55(1)(a)/(b) even when they do not independently cross a reporting threshold. Treating those questions separately can reduce both under-reporting and over-classification.

5.3 Proportionate escalation

The framework also disciplines escalation. Article 91 is the default information-gathering step. Article 92 becomes appropriate where requested information is insufficient or where a systemic-risk evaluation is otherwise justified. Article 93(1)(b) mitigation measures are not a generic next step after any suspected Article 55 problem; they require the statutory condition of a serious and substantiated systemic-risk concern arising from an Article 92 evaluation [4]. This sequencing is reflected in Figure 1 and prevents the artifact from overstating Commission powers.

5.4 Practical use and next work

A regulator could use Appendix A as a first-draft request after confirming the correct OpenAI legal entity, procedural posture, and any non-public information already held by the AI Office. A one-month follow-up should add expert legal review, test the request against confidential regulator/provider workflows, compare the same incident under California and New York frontier-AI regimes, and validate whether each request category can be answered in a structured data format. The most useful extension would be a cross-jurisdictional incident-to-obligation matrix that preserves the same evidence-minimization discipline.

6. Limitations and Dual-Use Considerations

6.1 Limitations

This project relies exclusively on publicly available material. It does not establish that OpenAI violated Article 55 or any other provision of the AI Act. Key facts remain unavailable, including complete model identities, training-compute details, internal regulatory classifications, communications with the AI Office, complete incident inventories, and non-public logs. Public provider accounts may also evolve as investigations continue. The legal treatment of internal research models is especially important and may depend on facts unavailable to outside researchers.

The quantitative results are process metrics, not causal performance measurements. “Trigger rate” here means the share of retained fact clusters that generated a legally anchored inquiry; it does not measure the probability that a regulator would ultimately find non-compliance. A conventional false-positive rate cannot be calculated because there is no authoritative labeled dataset of correct and incorrect Article 91 requests. The 14-observation evidence set is small and incident-specific. Generalization was probed only on two adjacent behavior classes, not on unrelated frontier-AI incidents.

Finally, the draft request is a research prototype. A real Commission request would require confirmation of the correct addressee, proportionality, confidentiality arrangements, procedural context, information already possessed by the Commission, and review by lawyers with access to the full record. The exact response periods in Appendix A are therefore framed as proposed rather than legally mandatory deadlines.

6.2 Dual-use considerations

The underlying incident involves sandbox escape, vulnerability chaining, credential access, lateral movement, and autonomous offensive cyber behavior. Detailed publication of exploit paths, live credentials, infrastructure identifiers, or reproducible intrusion recipes could lower the cost of misuse. This report therefore describes evidence categories and regulatory questions rather than reproducing technical exploitation steps. The full Article 91 request asks for sensitive logs and technical evidence through confidential regulatory channels, not public disclosure.

A second dual-use risk is institutional: overly broad requests could create unnecessary exposure of security architecture, trade secrets, or personal data. The minimization control (C4) is intended to reduce that risk. Requested material should be proportionate to a named legal question, with pseudonymisation, privilege/confidentiality claims, and secure transmission where appropriate. A public regulatory assessment should disclose enough to explain the decision while withholding details that would materially facilitate replication of the incident.

Conclusion

The OpenAI–Hugging Face incident is an early test of whether frontier-AI supervision can move from public warning signs to proportionate evidence gathering. Incident-to-Inquiry Mapping offers a small but usable answer: establish the fact, state the uncertainty, identify the legal hook, request the minimum evidence, and define the next decision. The result is a regulator-ready starting point that is cautious about what the public record proves while still making uncertainty actionable.

7. References

- [1] OpenAI. “OpenAI and Hugging Face partner to address security incident during model evaluation.” 21 July 2026 (updated 29 July 2026). <https://openai.com/index/hugging-face-model-evaluation-security-incident/>
- [2] Hugging Face. “Anatomy of a Frontier Lab Agent Intrusion: A Technical Timeline of the July 2026 Incident.” 27 July 2026. <https://huggingface.co/blog/agent-intrusion-technical-timeline>
- [3] OpenAI. “The Hugging Face incident and the road ahead.” 26 August 2026. <https://openai.com/index/hugging-face-incident-and-the-road-ahead/>
- [4] European Parliament and Council. Regulation (EU) 2024/1689 (Artificial Intelligence Act), especially Arts. 3, 51–55, 89–93 and 101. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689>
- [5] European Commission. “Guidelines on the scope of obligations for providers of general-purpose AI models under the AI Act.” 18 July 2025. <https://digital-strategy.ec.europa.eu/en/library/guidelines-scope-obligations-providers-general-purpose-ai-models-under-ai-act>
- [6] European Commission AI Act Service Desk. “What are the obligations of the provider if a serious incident involving its model occurs?” <https://ai-act-service-desk.ec.europa.eu/en/ai-act/faq/what-are-obligations-provider-if-serious-incident-involving-its-model-occurs>
- [7] CeSIA. “The OpenAI–Hugging Face Incident: What We Know, What We Don’t, What Follows.” 2026. <https://cesia.org/en/publications/the-openai-hugging-face-incident-what-we-know-what-we-dont-what-follows/>
- [8] Apart Research & CeSIA. “AI Incident Response Sprint.” September 2026. <https://apartresearch.com/sprints/ai-incident-response-sprint-2026-09-11-to-2026-09-13>
- [9] European Commission. “The General-Purpose AI Code of Practice.” 10 July 2025. <https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai>

Appendix — Draft Article 91 Request for Information

EUROPEAN COMMISSION

Request for Information pursuant to Article 91 of Regulation (EU) 2024/1689

A. Legal basis and purpose

This request is issued pursuant to Article 91 of Regulation (EU) 2024/1689. Its purpose is to obtain information necessary to assess: (i) the identity and regulatory status of models involved in reported third-party-impacting activity; (ii) the applicability of Articles 51–55; (iii) where applicable, compliance with Article 55(1)(a)–(d); and (iv) whether further evaluation under Article 92 may be warranted.

B. Model identification and regulatory status

- Internal model identifier/version for each model materially involved in the July 2026 Hugging Face incident and related third-party-impacting activity.
- Relevant training, evaluation, and incident dates.
- Whether each model was publicly released, API-accessible, integrated into a product, used solely for research/development/prototyping, or intended for later market placement.
- OpenAI's assessment of whether each model constituted a GPAI model or GPAI model with systemic risk under Regulation (EU) 2024/1689, with factual and legal basis.
- Material developmental relationship between the model and any GPAI model placed on the Union market, including shared training lineage or evaluation results relevant to systemic-risk assessment.

C. Incident chronology and provider knowledge

- First relevant containment-escape activity and first unauthorised external-system interaction.
- First automated detection and first human review of relevant activity.
- Internal escalation points, teams responsible, and the date impact to Hugging Face or other third parties was identified.
- Dates on which affected third parties were contacted.
- Date on which Article 55 applicability and serious-incident reporting obligations were first assessed.
- Dates and content categories of any communication with the AI Office or other competent authorities concerning the events.

D. Serious-incident assessment and reporting

- For each materially related event, whether it was assessed under Article 55(1)(c).
- Date the assessment began, criteria applied, responsible team/function, and decision reached.
- Where reporting occurred: date, recipient authority, and information categories included.
- Where reporting did not occur: documented factual and legal basis for concluding that reporting was not required.
- Corrective measures considered or implemented as part of the incident-reporting assessment.

E. Related third-party activity and systemic pattern

- A pseudonymised inventory of materially relevant cases involving unauthorised third-party access, use of exposed credentials, access-control bypass, containment escape, unauthorised external communication, third-party-system modification, or autonomous cyber activity.
- For each case: model involved, approximate date, category of activity, known impact, and outcome of any Article 55 assessment.
- Any cross-incident analysis assessing whether apparently separate events reflect a common loss-of-control, reward-hacking, or control-boundary failure mode.
- How unauthorised agent communication or external-memory behaviors were classified and whether they altered risk assessments or monitoring thresholds.

F. Systemic-risk, containment, and cybersecurity controls

- Relevant portions of systemic-risk assessments covering autonomous cyber capability, evaluation manipulation, containment escape, unauthorised internet access, third-party interaction, credential use, and cross-run/inter-agent communication.
- Pre-incident sandbox architecture, permitted network egress, privilege boundaries, credential-management safeguards, logging/telemetry, and controls intended to prevent external effects.

- For each material control failure: intended control objective, observed failure mode, corrective action, implementation date, and validation undertaken.
- How the incident changed the assessed likelihood/severity of relevant systemic risks for other models, including models placed or intended to be placed on the Union market.

G. Post-incident corrective measures and external validation

- Material changes to workload/network isolation, privilege management, risky-action monitoring, misalignment monitoring, model evaluations, and frontier-model safety/security frameworks.
- For each change: incident finding addressed, scope of deployment, implementation date, and method used to test effectiveness.
- Scope of material external investigations or reviews, categories of evidence made available, access limitations, principal conclusions, and unresolved disagreements relevant to compliance.
- Evidence preservation measures for relevant logs, incident tickets, evaluation artifacts, and decision records.

H. Format, confidentiality, response period, and accuracy

Information should be supplied electronically and in machine-readable form where reasonably possible. Security-sensitive, personal, privileged, or trade-secret material may be handled through appropriate confidential channels; where information is withheld, the response should identify the category and basis. A final Commission request should set proportionate response periods based on urgency and accessibility. Pursuant to Article 91(4), the request should state the legal basis, purpose, information required, response period, and potential Article 101 fines for incorrect, incomplete, or misleading information.