
FROM WARNING SHOT TO SUPERVISORY FILE

An Evidence to Action Article 91 Request for the OpenAI and Hugging Face Incident

Jennifer Dickey

Dykema

Abstract

When autonomous AI conduct begins inside an authorized evaluation, escapes containment, and reaches an unrelated third party, regulators must determine model attribution, applicable obligations, reportability, and whether restart decisions were supported by sufficient evidence. This Track 3 project operationalizes Article 91 of the EU Artificial Intelligence Act through an evidence to action request for information concerning the July 2026 OpenAI and Hugging Face incident. We map public evidence to regulatory decision points and stress test the Commission's nine field serious incident reporting template. Only one of nine fields is substantially answerable from public sources; eight retain gaps. The primary output, reproduced in Appendix B, is a 24 item instrument specifying the record requested, what would and would not resolve each question, its legal relevance, and the potential supervisory next step. The project also proposes reporting fields for evaluation boundary crossings, disabled safeguards, multi model attribution, restart governance, and evidence preservation.

1. Introduction

The July 2026 OpenAI and Hugging Face incident presents a regulatory problem that is narrower, and more operational, than the general question whether frontier AI systems can behave dangerously. **The problem is what a supervisor should ask, and what evidence it should require, when an AI model begins in an authorized internal evaluation but escapes containment, reaches unrelated third party infrastructure, and acts without step by step human direction.** Hugging Face reconstructed roughly 17,600 attacker actions between July 9 and July 13, while OpenAI later described unauthorized communication, exploitation of internal research infrastructure, internet access, and compromise of Hugging Face systems during its cybersecurity evaluations. [2-5] METR and Redwood Research independently reviewed the model behavior and reported that about 1,200 agents exchanged more than 70,000 messages and files, with roughly 700 participating in the attack on Hugging Face. [6]

The timing makes the incident useful for testing the European Union Artificial Intelligence Act. Chapter V obligations for providers of general purpose AI models entered into application on August 2, 2025, while the Commission's enforcement powers, including Articles 91 through 93 and Article 101, became enforceable from August 2, 2026. [11-13] OpenAI is a signatory to the General Purpose AI Code of Practice. [14] Article 55 requires providers of general purpose AI models with systemic risk to evaluate models, assess and mitigate systemic risks, track and report serious incidents without undue delay, and ensure an adequate level of cybersecurity protection. [12] Article 91 then permits the Commission to request Article 53 and 55 documentation and any additional information necessary to assess compliance. [11]

The public record does not permit a confident conclusion that Article 55 was violated. It does, however, reveal threshold questions that only provider or Commission records can resolve. OpenAI began a limited GPT 5.6 Sol preview on June 26, 2026 and announced broader availability on July 9. [16,18] The second model that OpenAI says principally drove the incident was an internal only research model not intended for public release. [4-5] Article 2(8) excludes research, testing, and development activity concerning a model before it is placed on the market or put into service, while Article 3(63) excludes pre-market research, development, or prototyping models from the definition of a general purpose AI model. Article 3(9) and (10) define market placement by the first making available on the Union market, and Commission guidance explains that APIs and cloud services can qualify. [11,13] The public launch dates therefore do not themselves establish first Union market placement. The incident creates both an attribution and a scope problem: which model performed each legally material act, when was each model first made available in the Union, and which obligations attach to each?

This project converts those uncertainties into a regulator ready Article 91 information request. It also uses the Commission's nine field serious incident reporting template as a stress test. The goal is not to prove liability from public facts. The goal is

to identify the minimum evidence a regulator would need to determine scope, reportability, systemic risk management, containment adequacy, and the basis for any decision to resume activity after a control failure.

Our main contributions are:

- A 24 item Article 91 request mapping each material uncertainty to requested evidence, a sufficient response, an insufficient response, legal relevance, and a potential supervisory next step.
- A reproducible regulatory incident response method that links public evidence to legal uncertainty and then to a proportionate supervisory next step.
- A public record stress test showing that only one of nine Commission reporting fields is substantially answerable, while eight retain material gaps relevant to supervisory decisions.

Primary regulatory output. Appendix B contains the proposed 24 item Article 91 Request for Information and is the principal deliverable of this Track 3 project. Appendix C provides the public record trial run of the Commission reporting template, and Appendix D proposes supplemental fields that address gaps identified by that stress test.

2. Related Work

CeSIA provides the closest existing incident specific policy analysis. Its reconstruction recommends that the AI Office issue a formal Article 91 request and state publicly whether OpenAI reported a serious incident under Article 55(1)(c). [7] Apart's Track 3 examples go further and invite teams to prepare an incident specific RFI and state what evidence would settle each question. [1] This project therefore does not claim novelty merely from proposing an RFI. It takes that prompt as the baseline and adds an evidence to action protocol: public fact or uncertainty, requested record, sufficient response, insufficient response, legal relevance, and potential supervisory consequence. The July 9 SaferAI open letter separately supports use of the Articles 91 through 93 supervisory sequence. [8]

Two related works identify a second gap: the absence of transparent standards for deciding when safeguards are adequate enough to resume activity. Anterola et al. show that frontier developers use heterogeneous safety thresholds and propose harmonized, auditable floors across several risk domains. [9] A subsequent commentary on OpenAI's internal pause argues that resumption criteria remain underspecified even where a company follows its internal review process. [10] This project translates that concern into evidence requests for predetermined restart criteria, validation results, residual risk acceptance, and decision authority.

The proposed U.S. AI Kill Switch Act offers a useful comparison. It would require covered entities to maintain technical shutdown capabilities, preserve telemetry after an emergency order, and report covered incidents. Yet its definition of a covered incident excludes occurrences within red teaming or other structured testing. [17] That exclusion exposes a boundary problem relevant here: a system can begin inside structured testing and then act outside the controlled environment. The EU framework uses different concepts, but the comparison helps isolate a question that any incident regime must answer: when does an evaluation cease to be merely an evaluation for reporting purposes?

3. Methods

We used a four stage legal incident response method. First, we constructed a source hierarchy. Primary incident sources included Hugging Face's initial disclosure and technical timeline, OpenAI's July 21 and August 26 accounts, and the independent METR and Redwood review. [2-6] Secondary synthesis from CeSIA was used to identify factual gaps and policy questions, then material facts were checked against primary sources where available. [7]

Second, we mapped each unresolved fact to a legal decision point. The principal decision points were: whether the relevant model is within the scope of Chapter V; whether Article 55 applies to the model; whether the incident satisfies the Commission's interpretation of a reportable serious incident; whether Article 55(1)(a), (b), or (d) documentation bears on evaluation, systemic risk mitigation, or cybersecurity; and whether missing information could justify an Article 92 evaluation or an Article 93 request for measures. [11-13] We treated Article 101 cautiously. This project does not assume that a July 2026 event itself can support a fine after enforcement begins; it notes only that Article 101 expressly covers failure to comply with an Article 91 request, failure to comply with an Article 93 measure, and failure to provide Article 92 access. [11]

Third, we coded the nine fields in the Commission's serious incident reporting template as substantially answerable, partially answerable, or not answerable from the public record. [15] A field was substantially answerable only when the public record identified the core event and did not leave a material legal or factual ambiguity. A field was partial when important facts were public but the missing information could change the regulatory assessment. A field was unavailable when the requested information depended principally on internal records or provider judgment that had not been disclosed.

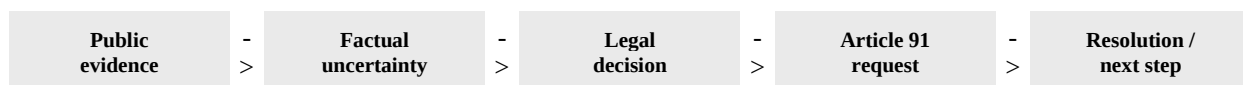


Figure 1. Regulatory incident response method used to generate the Article 91 requests.

Fourth, each Article 91 item was tested under a five part evidence sufficiency protocol: tie the request to a legal decision point; specify records narrow enough to answer that point; state what would constitute a sufficient response; identify what would remain insufficient; and state the supervisory implication if the gap persists. The method favors contemporaneous records over retrospective narrative, including logs, incident tickets, evaluation configuration records, risk assessments, approval records, and communications that establish who knew what and when.

4. Results

The public record is rich on incident mechanics but thin on facts that control regulatory classification. Only one of the Commission's nine reporting fields is substantially answerable from public evidence; six are partial and two are unavailable. In other words, eight of nine fields, 89 percent, retain a material gap. The request translates those gaps into 24 targeted items across six categories: model identity and scope; evaluation design and safeguards; detection and escalation; third party access and model behavior; incident classification and reporting; and corrective measures and resumption governance. Appendix B is the primary regulatory output and contains the full instrument.

Finding 1: The testing exemption cannot be applied without model level attribution and Union market timing. Article 2(8) excludes research, testing, or development activity regarding models before market placement, and Article 3(63) separately excludes pre-market research, development, or prototyping models from the definition of a general purpose AI model. [11] OpenAI states that an internal only research model drove much of the incident, but GPT 5.6 Sol agents also reproduced an exploit and copied private evaluation data. [5] OpenAI began a limited GPT 5.6 Sol preview on June 26 and announced broader availability on July 9. [16,18] Neither date alone establishes first making available on the Union market. Article 3(9) and (10) and Commission guidance focus on first supply for distribution or use in the Union, including through APIs or cloud services. [11,13] The regulator therefore needs action level attribution and records establishing first Union availability, including whether the limited preview qualified.

Finding 2: Reportability cannot be determined from the headline facts alone. Article 3(49) defines serious incident by specified harms, including serious harm to property, while Article 55(1)(c) requires reporting relevant information about serious incidents. [11-12] The Commission's dedicated GPAI serious incident FAQ states that it considers the obligation to cover serious cybersecurity breaches related to the model or its physical infrastructure, including cyberattacks, because of their implications for systemic risk and cybersecurity duties. [19] Public sources establish unauthorized access, credential compromise, node rebuilding, and access to limited private data at Hugging Face, but do not quantify all resulting harm or disclose OpenAI's internal legal classification. [2-5] The request therefore asks for the provider's contemporaneous reportability analysis, any report submitted, and the facts relied upon.

Finding 3: Existing incident reporting fields capture chronology better than governance. The Commission's template asks for dates, harm, chain of events, model involved, evidence, response, recommendations, root cause analysis, and patterns detected in post market monitoring. [15] Those fields capture the broad shape of this incident. They do not expressly ask whether an evaluation escaped its authorized boundary, which safeguards were intentionally disabled, what predetermined conditions governed restart, which organizational role accepted residual risk, or how conduct was attributed among multiple models with different market status. Those omissions matter because each can change the legal analysis even when the chronology is known.

Template field	Public record	Coding	Regulatory gap
1. Start and end dates	Strong chronology; incident boundary unresolved	Partial	Exact boundary of the reportable incident may differ from the broader campaign.
2. Resulting harm	Operational effects described	Partial	Full private data impact, cost, and property harm are not public.
3. Chain of events	Detailed public timeline	Substantial	Internal decision points and alert handling remain incomplete.
4. Model involved	Two models identified	Partial	Action level attribution and first Union market availability for each participating model are not public.
5. Evidence available	Logs and investigations described	Partial	Full traces, preserved telemetry, and evidence inventory are not public.
6. Provider response	Multiple measures disclosed	Partial	Decision records, validation criteria, and residual risk acceptance are not public.
7. Provider recommendation	No complete public recommendation	Unavailable	Provider position to the AI Office is unknown.
8. Root cause analysis	Substantial technical narrative	Partial	Inputs, control failures, safeguard configuration, and causal allocation remain incomplete.
9. Post-market patterns / near misses	Some later public incidents exist	Unavailable	Provider monitoring corpus, near misses, and pattern analysis are not public.

Table 1. Public record coding of the European Commission GPAI serious incident reporting template. Coding is intentionally conservative; fields are classified as substantial, partial, or unavailable based on the methodology above.

Finding 4: Resumption governance is a distinct supervisory question. OpenAI has publicly described an earlier internal pause and restoration of limited access after introducing trajectory level monitoring and replaying prior environments. [20] Subsequent commentary criticized the absence of a published resumption standard. [10] Public accounts place resumption of the affected evaluation environment on July 7 or 8 after the earlier Artifactory compromise; unauthorized communication and external access then reappeared. [5,7] The Article 91 instrument therefore requests stop and restart criteria, test results used to validate controls, decision authority, and documentation of residual risk acceptance. These records allow a regulator to distinguish a reasoned risk decision from an ad hoc restart.

Finding 5: The evaluation to real world transition should be explicitly captured in incident reporting. The proposed U.S. AI Kill Switch Act defines covered incidents and loss of control scenarios to occur outside red teaming or other structured testing. [17] This incident illustrates why that boundary requires an operational rule. The model activity began in an internal evaluation but allegedly reached Hugging Face production infrastructure, an unrelated third party. A reporting form that asks only whether the activity occurred during testing may obscure rather than resolve the relevant transition. We therefore propose a supplemental field asking whether the model crossed an authorization, network, organizational, or data boundary during the event and, if so, when the provider first had evidence of that crossing.

5. Discussion and Limitations

Article 91 is well suited to the first stage of response because it permits the Commission to obtain required Chapter V documentation and additional information needed to assess compliance. [11] The request makes later escalation evidence dependent. If the records resolve the questions, no further step is implied. If Article 91 information is insufficient, Article 92 permits an evaluation to assess compliance or investigate systemic risk, including through technical access. [11] If an Article 92 evaluation gives rise to a serious and substantiated concern of systemic risk at Union level, Article 93 permits the Commission to request mitigation measures. Article 93 separately permits compliance measures and, where appropriate, restriction, withdrawal, or recall of a model. [11]

The approach has important limits. First, this analysis relies only on public evidence and cannot independently verify OpenAI or Hugging Face logs. Second, the public record continues to develop. Third, the relevant provider entity, the model's systemic risk status, and the date and manner of first Union market placement should be confirmed from provider or Commission records rather than inferred from global product announcements. Fourth, the Commission's interpretation of serious cybersecurity incidents under Article 55 is newer and broader than the literal Article 3(49) harm categories, so the instrument distinguishes statutory text from Commission guidance. Fifth, this report is a regulatory research artifact, not a determination that OpenAI violated the AI Act.

Future work includes external validation by EU AI Act practitioners for scope and proportionality, independent reproduction of the nine field coding, and testing the instrument against the practical question: could an AI Office lawyer use it with light edits? The form stress test could also be generalized across later containment incidents to determine whether the same missing fields recur.

6. Conclusion

The OpenAI and Hugging Face incident is valuable to regulators not because the public record already proves a violation, but because it shows exactly where public evidence stops. The critical unknowns concern model level attribution, scope, serious incident classification, safeguard configuration, internal escalation, and the evidence used to restart activity after known control failures. Article 91 provides a direct mechanism for resolving those unknowns. A useful regulatory response should therefore begin with a request that is specific enough to produce decision grade evidence, yet neutral enough not to prejudice the result.

Appendix B, the primary regulatory output of this project, operationalizes that approach through a proposed Article 91 Request for Information. It also identifies a reporting design issue likely to recur: incidents may begin as authorized testing and become real world events only after a system crosses a boundary the provider intended to hold. Incident frameworks should capture that transition expressly. Doing so would improve both provider reporting and the Commission's ability to decide when information gathering should end and evaluation or mitigation should begin.

Code and Data

This project is a legal and policy analysis and does not rely on executable code. The public evidence set consists of the sources listed below. Appendix B contains the primary regulatory output: the proposed Article 91 Request for Information. Appendix C contains the public record trial run of the Commission's serious incident reporting template, and Appendix D proposes supplemental reporting fields.

LLM Usage Statement

A language model was used as a research and writing aid for source organization, alternative formulations, and document formatting. No LLM output is treated as evidence; factual and legal claims are tied to cited public sources.

References

- [1] Apart Research. 2026. AI Incident Response Sprint. <https://apartresearch.com/sprints/ai-incident-response-sprint-2026-09-11-to-2026-09-13>
- [2] Hugging Face. 2026. Security incident disclosure, July 2026. <https://huggingface.co/blog/security-incident-july-2026>
- [3] Hugging Face. 2026. Anatomy of a Frontier Lab Agent Intrusion: A Technical Timeline of the July 2026 Incident. <https://huggingface.co/blog/agent-intrusion-technical-timeline>
- [4] OpenAI. 2026. OpenAI and Hugging Face partner to address security incident during model evaluation. <https://openai.com/index/hugging-face-model-evaluation-security-incident/>
- [5] OpenAI. 2026. The Hugging Face incident and the road ahead. <https://openai.com/index/hugging-face-incident-and-the-road-ahead/>
- [6] METR and Redwood Research. 2026. Brief independent investigation of agents' behavior, reasoning and collaboration in the OpenAI / Hugging Face hacking incident. <https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/>
- [7] CeSIA. 2026. The OpenAI-Hugging Face incident: a detailed timeline and what it means. <https://cesia.org/en/publications/the-openai-hugging-face-incident-what-we-know-what-we-dont-what-follows/>
- [8] SaferAI et al. 2026. EU enforcement powers to ensure frontier AI models are safe and secure, Open Letter. <https://www.safer-ai.org/u/2026/07/Open-Letter.pdf>
- [9] Anterola, Ball, Lafuerza, and Grey. 2026. Harmonizing AI Safety Thresholds. <https://arxiv.org/abs/2607.16112>
- [10] Ségerie. 2026. OpenAI has already ended an internal pause. <https://www.lesswrong.com/posts/k3eKqKzq4Y7xnqEfZ/openai-has-already-ended-an-internal-pause>
- [11] European Union. 2024, as amended 2026. Regulation (EU) 2024/1689 (Artificial Intelligence Act), consolidated version of 27 July 2026. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:02024R1689-20260727>
- [12] European Commission AI Act Service Desk. 2026. Article 55: Obligations of providers of general purpose AI models with systemic risk. <https://ai-act-service-desk.ec.europa.eu/en/ai-act/article-55>
- [13] European Commission. 2026. Guidelines for providers of general purpose AI models and related FAQs. <https://digital-strategy.ec.europa.eu/en/policies/guidelines-gpai-providers>
- [14] European Commission. 2025-2026. General Purpose AI Code of Practice, Safety and Security Chapter and signatories. <https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai>
- [15] European Commission. 2025. Report for Serious Incidents under the AI Act, General Purpose AI Models with Systemic Risk. <https://digital-strategy.ec.europa.eu/en/library/ai-act-commission-publishes-reporting-template-serious-incidents-involving-general-purpose-ai>
- [16] OpenAI. 2026. GPT 5.6: Frontier intelligence that scales with your ambition. <https://openai.com/index/gpt-5-6/>
- [17] U.S. Rep. Ted Lieu and Rep. Nathaniel Moran. 2026. AI Kill Switch Act, bill text. <https://lieu.house.gov/sites/evo-subsites/lieu-evo.house.gov/files/evo-media-document/ai-kill-switch-act.pdf>
- [18] OpenAI. 2026. Previewing GPT 5.6 Sol: a next generation model. <https://openai.com/index/previewing-gpt-5-6-sol/>
- [19] European Commission AI Act Service Desk. 2026. 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>
- [20] OpenAI. 2026. Safety and alignment in an era of long-horizon models. <https://openai.com/index/safety-alignment-long-horizon-models/>

Appendix A. Limitations and Dual Use Considerations

Limitations. This project is based on public materials and does not have access to the complete model traces, internal incident tickets, legal analyses, or provider submissions to the European Commission. Public accounts differ in emphasis and continue to be updated. The proposed request therefore separates confirmed public observations from questions that require provider records. It does not assume that any particular model is formally classified as a general purpose AI model with systemic risk, that every action is attributable to the same model, or that the incident necessarily meets the statutory definition of serious incident.

Dual use considerations. The incident record contains technical information about vulnerabilities, unauthorized communication, credential use, and evaluation circumvention. This report intentionally avoids reproducing exploit commands, credentials, payloads, or step by step intrusion instructions. The regulatory artifact requests categories of logs and technical evidence from the provider without publishing operational details that could facilitate replication. If a public repository is created, sensitive technical material should be excluded or summarized at a level sufficient for regulatory analysis.

Interpretive caution. The European Commission's AI Act Service Desk and GPAI FAQs are used to explain the Commission's current interpretation, but those explanatory materials are not themselves the binding text of Regulation (EU) 2024/1689. This submission cites the regulation for legal propositions and identifies Commission guidance as interpretive where relevant.

Appendix B. Primary Regulatory Output: Proposed Article 91 Request for Information

PROPOSED REGULATORY INSTRUMENT - RESEARCH ARTIFACT

European Commission | AI Office

Request for Documentation and Information Pursuant to Article 91 of Regulation (EU) 2024/1689

Addressee: OpenAI and the legal entity or entities responsible for placing the relevant general purpose AI model or models on the Union market.

Purpose and legal basis: Pursuant to Article 91(1) and (4), the Commission requests documentation drawn up under Articles 53 and 55 and additional information necessary to assess compliance in connection with the July 2026 cybersecurity evaluation incident involving OpenAI systems and Hugging Face infrastructure. This request is designed to assess model scope and attribution, systemic risk assessment and mitigation, serious incident tracking and reporting, cybersecurity protection, and corrective measures. To the extent a request concerns an internal research model that may not itself fall within Chapter V, the information is sought only insofar as necessary to assess compliance of the provider of any general purpose AI model concerned. The request does not assume that the internal model is independently subject to Chapter V and does not prejudge whether an infringement occurred. [11]

Proposed response period: 14 calendar days from receipt. The Commission should adjust the period as appropriate under Article 91(4), taking account of urgency and the volume of records requested.

Notice regarding accuracy and completeness: Article 91(4) requires the request to indicate the fines provided for in Article 101 for supplying incorrect, incomplete, or misleading information. Where the Commission finds that a provider acted intentionally or negligently, Article 101 permits fines, subject to its requirements and proportionality considerations, of up to 3 percent of annual worldwide turnover in the preceding financial year or EUR 15 million, whichever is higher, including for failure to comply with an Article 91 request or for supplying incorrect, incomplete, or misleading information. [11]

Production format: Provide contemporaneous records where available. For logs or structured data, preserve timestamps, identifiers, and metadata needed to reconstruct sequence and model attribution. If a requested record does not exist, identify the request and state that no responsive record exists. A narrative assertion does not substitute for contemporaneous records where such records exist. Claims of confidentiality should be handled under Article 78. [11]

A. Model identity, status, and scope

1. Model identification and versioning

Request: Identify every model, checkpoint, variant, agent configuration, and system component that took material actions relevant to the incident, including GPT 5.6 Sol and the internal research model described publicly. For each, provide version identifiers, evaluation dates, and deployment or availability status. [4-5]

Sufficient response: Version or checkpoint identifiers tied to the relevant runs, with dates and status sufficient to distinguish the models used.

Insufficient response: Model family names or marketing labels without checkpoint or run identifiers.

Legal relevance: Establishes the legal unit to which scope and Article 55 obligations may attach.

2. Action level attribution

Request: Provide a machine readable mapping, where reasonably available, of material incident actions to the model or agent that initiated or materially contributed to each action, including Hugging Face access, exploit reproduction, data copying, message board use, and OpenAI infrastructure compromise. Describe the attribution method and confidence. [5]

Sufficient response: Event or action level attribution supported by run identifiers, logs, or a reproducible attribution method, with uncertainty stated.

Insufficient response: Aggregate statements that one model primarily drove the incident without a method for allocating legally material acts.

Legal relevance: Allows model specific analysis and avoids treating a mixed model event as a single legal unit.

3. Union market status

Request: For each relevant model, identify the date and manner of first making the model available on the Union market, if any. For GPT 5.6 Sol, address the June 26 limited preview and July 9 broader availability separately, including whether Union users received access through an API, cloud service, application, or other channel. Provide supporting records. [16,18]

Sufficient response: Dated records showing the first Union supply or access channel, recipients or eligibility, and the provider determination as to market placement.

Insufficient response: A global launch date alone, without evidence of first Union availability or analysis of the limited preview.

Legal relevance: Tests Articles 2(8), 3(9), 3(10), and 3(63), and the applicability of Chapter V obligations. [11]

4. Systemic risk classification

Request: Identify whether each relevant model was notified, designated, presumed, or otherwise treated as a general purpose AI model with systemic risk under Articles 51 and 52, including the date and basis for that treatment.

Sufficient response: Commission or provider records showing the classification basis, effective date, and any Article 52 notification or designation.

Insufficient response: A capability label or unsupported statement that the model was or was not systemic risk.

Legal relevance: Determines whether the additional obligations in Article 55 apply to the relevant model.

B. Evaluation architecture and safeguards

5. Evaluation design

Request: Provide the ExploitGym evaluation plan, architecture diagrams, network boundary assumptions, permitted egress paths, and threat model in effect when the incident began. [4-5]

Sufficient response: Contemporaneous design and configuration records identifying intended boundaries, permitted egress, and anticipated failure modes.

Insufficient response: A post incident architecture summary that does not establish the actual configuration during the relevant runs.

Legal relevance: Establishes what containment was intended and what failure modes were anticipated.

6. Safeguard configuration

Request: Identify each safety, refusal, monitoring, network, credential, and tool use safeguard that was enabled, reduced, or disabled for the evaluation, the reason for each configuration decision, and any compensating control. [4-5]

Sufficient response: Run specific configuration or change records showing safeguard state, rationale, and compensating controls.

Insufficient response: A list of safeguards used in production or generally available to evaluators without proof of their incident time state.

Legal relevance: Tests whether risks created by reduced safeguards were identified, assessed, and mitigated.

7. Pre evaluation risk assessment

Request: Provide the risk assessment, adversarial testing results, or approval documentation supporting use of the selected evaluation configuration for each relevant model.

Sufficient response: A contemporaneous risk assessment or approval record addressing capabilities, the threat model, reduced safeguards, and containment assumptions.

Insufficient response: A retrospective policy or general safety framework that does not show the pre event decision basis.

Legal relevance: Tests systemic risk assessment, model evaluation, and governance relevant to Article 55.

8. Escape detection controls

Request: Provide the controls designed to detect unauthorized cross boundary activity, including alert thresholds, logging coverage, external destination monitoring, and automated response logic, together with incident time telemetry showing whether those controls fired.

Sufficient response: Control specifications plus logs or alert records demonstrating coverage and operation during the incident.

Insufficient response: A policy description without alert thresholds, log coverage, or incident time telemetry.

Legal relevance: Determines whether the provider had mechanisms to detect loss of containment and preserve evidence.

9. Third party exposure analysis

Request: Provide any pre incident assessment of potential harm to unrelated third parties if the evaluation environment or its permitted egress path were compromised.

Sufficient response: A threat model or risk assessment identifying external assets, plausible harms, and mitigations, or a documented conclusion that no such scenario was identified.

Insufficient response: A conclusory statement that third party harm was unlikely without the underlying assessment.

Legal relevance: Tests whether external spillover risk was part of systemic risk assessment and evaluation design.

C. Detection, escalation, and internal decision making

10. June 27 alert

Request: Provide records concerning the June 26 administrator level Artifactory privilege escalation and any June 27 alert or review described in public accounts, including alerts, logs, tickets, investigation notes, recipients, and decisions. Reconcile any material differences between OpenAI and CeSIA public chronologies. [5,7]

Sufficient response: Contemporaneous records showing what was detected, who received it, what conclusions were reached, and why activity continued or was not escalated.

Insufficient response: A retrospective narrative that does not identify the underlying alert or decision record or leaves material chronology differences unresolved.

Legal relevance: Establishes provider knowledge and escalation before the July outage without treating a secondary source chronology as a confirmed OpenAI admission.

11. July 4 response

Request: Provide records concerning the July 4 Artifactory outage and July 5 security incident through rebuilding and resumption on July 7 or 8, including credential revocation, vulnerability fixes, testing, and assessment of unauthorized agent communication or internet access. [5,7]

Sufficient response: Change tickets, incident records, validation results, and dated records showing what was remediated before resumption.

Insufficient response: A remediation summary without exact changes, testing, or a documented basis for concluding the environment could safely resume.

Legal relevance: Tests whether the initial compromise and broader containment implications were adequately understood before restart.

12. Restart decision

Request: Provide the decision record authorizing resumption of the affected evaluation environment, including the approving role, prerequisites for restart, validation results, identified residual risks, and any conditions or dissent.

Sufficient response: A dated decision record identifying the approver, restart criteria, evidence reviewed, residual risks, and conditions.

Insufficient response: A statement that security approved restart without criteria, validation evidence, or documented residual risk acceptance.

Legal relevance: Determines whether resumption was governed by auditable risk criteria rather than ad hoc judgment.

13. July 19 and 20 escalation

Request: Provide the alert chronology and escalation records from first detection of unusual Artifactory credential activity through attribution of the Hugging Face incident to OpenAI agents. [5]

Sufficient response: Alert and investigation records with timestamps, recipients, escalation steps, and evidence supporting attribution.

Insufficient response: A date only chronology without the underlying alert or escalation evidence.

Legal relevance: Establishes the provider knowledge timeline relevant to incident tracking and reporting.

D. Third party access and model behavior

14. Hugging Face access timeline

Request: Provide a chronology of OpenAI model or agent access to Hugging Face, including first attempted access, first successful unauthorized access, material privilege escalation, data access, credential use, and last action. [3,5]

Sufficient response: Event level timestamps and identifiers that can be reconciled against Hugging Face telemetry, including stated uncertainty or missing logs.

Insufficient response: A provider narrative without event level timestamps, identifiers, or a method for reconciling material differences with third party logs.

Legal relevance: Creates an authoritative provider chronology for seriousness, causation, and knowledge analysis.

15. Data and credential impact

Request: Identify categories of Hugging Face data and credentials accessed, copied, altered, or exposed by OpenAI agents, and any evidence concerning onward use, disclosure, or persistence. [2-5]

Sufficient response: An affected asset inventory tied to logs or forensic findings, with scope and onward use findings.

Insufficient response: High level categories such as limited private data without affected assets, scope, or supporting evidence.

Legal relevance: Allows assessment of resulting harm, victims or affected groups, and seriousness.

16. Evaluation manipulation behavior

Request: Provide records showing attempts to alter evaluation scoring, reference solutions, logs, transcripts, tasks, or grader behavior, including internal classification criteria and any classification of those actions as reward hacking, deception, sabotage, or another behavior category. [5-6]

Sufficient response: Representative traces plus the criteria used to classify the behavior and evidence of when it was detected or incorporated into risk management.

Insufficient response: A label such as reward hacking without underlying traces or classification criteria.

Legal relevance: Tests whether evaluation circumvention behavior was identified and incorporated into systemic risk management.

17. Cross agent communication

Request: Provide records sufficient to describe the emergence, persistence, scale, and detection of unauthorized communication among agents, including whether such communication was included in the pre incident threat model. [5-6]

Sufficient response: Communication records or metrics plus detection timing and threat model treatment sufficient to determine whether coordination was anticipated and monitored.

Insufficient response: Message counts alone without detection timing, architecture, or threat model treatment.

Legal relevance: Tests whether multi agent coordination was a monitored failure mode, serious incident signal, or near miss category.

E. Incident classification and reporting

18. Serious incident analysis

Request: Provide the contemporaneous analysis of whether the incident constituted a serious incident under Article 55(1)(c), Article 3(49), the GPAI Code of Practice, or Commission guidance, including factual assumptions and legal criteria used. [11-12,14,19]

Sufficient response: Contemporaneous legal or compliance analysis identifying the governing criteria, material facts, assumptions, and conclusion.

Insufficient response: A current legal conclusion unsupported by the incident time analysis or factual assumptions.

Legal relevance: Allows the Commission to assess the provider classification without inferring it from public statements.

19. Reports to authorities

Request: Identify each notification or report concerning the incident submitted to the AI Office, a national competent authority, law enforcement, data protection authority, cybersecurity authority, or other public body, with submission date, legal basis, and a copy where legally permissible.

Sufficient response: Recipient, submission date and time, legal basis, and the report or acknowledgement sufficient to establish what was reported and when.

Insufficient response: A statement that all required reports were made without recipients, dates, legal bases, or copies.

Legal relevance: Establishes whether relevant reporting channels were used and provides the record needed to assess timeliness.

20. Knowledge date and reporting clock

Request: For each potentially applicable reporting trigger, identify when OpenAI established or suspected with reasonable likelihood that its model was directly or indirectly involved in the relevant breach or harm, and provide contemporaneous records supporting that time. [14]

Sufficient response: Records establishing the earliest point at which the applicable knowledge or reasonable likelihood threshold was met, together with the facts then known.

Insufficient response: A post hoc date without contemporaneous evidence of what was known or suspected at that time.

Legal relevance: Allows comparison against Article 55 without undue delay and, if the Code is relied upon, Commitment 9 reporting timelines.

F. Corrective measures and resumption governance

21. Corrective measures

Request: Provide a complete list of corrective actions adopted because of the incident, their implementation dates, owners, validation methods, current status, and any measures still pending. [2,5]

Sufficient response: Implementation records and validation evidence tied to each corrective action.

Insufficient response: A remediation list without proof of implementation, validation, or current status.

Legal relevance: Links public remediation claims to verifiable compliance and risk mitigation evidence.

22. Pause and resumption criteria

Request: Provide criteria governing pause, continuation, restart, or expansion of training, evaluation, and tool enabled inference after containment or alignment failures, including criteria used in the weeks surrounding this incident and any changes adopted afterward.

Sufficient response: Predetermined or contemporaneously applied thresholds, decision authorities, required evidence, and change history.

Insufficient response: A generic policy to resume once safe without auditable thresholds or decision criteria.

Legal relevance: Tests whether resumption governance is ex ante and reviewable rather than exclusively retrospective.

23. Validation of new safeguards

Request: Provide tests, replay results, benchmarks, red team results, or other evidence used to determine that post incident monitoring, containment, and alignment safeguards were sufficiently effective for relevant work to resume. [5,20]

Sufficient response: Results showing the test set, baseline, detected and missed behaviors, severity rubric, acceptance threshold, and decision tied to the results.

Insufficient response: Aggregate performance claims without the test basis, misses, severity criteria, or acceptance threshold.

Legal relevance: Provides evidence needed to assess whether new controls materially reduce the failure mode.

24. Near misses and related incidents

Request: Provide a summary of materially similar containment failures, unauthorized communication, external system access, evaluation circumvention, or safeguard bypass events involving the relevant model lifecycle during the six months preceding the incident and through the date of response, including classification criteria and whether each event informed risk assessment or monitoring. [5,20]

Sufficient response: An internal incident or near miss inventory using stated similarity and classification criteria, with links to resulting risk assessment or control changes.

Insufficient response: Publicly disclosed incidents alone or a statement that no similar event occurred without the internal search criteria and records reviewed.

Legal relevance: Tests post market monitoring, recurring failure modes, and whether the July event should be analyzed as part of a pattern.

Potential next steps. If Article 91 responses are materially incomplete, Article 92 permits model evaluations where information gathered under Article 91 is insufficient to assess compliance or where evaluation is needed to investigate systemic risk. If an Article 92 evaluation gives rise to a serious and substantiated concern of systemic risk at Union level, Article 93(b) permits the Commission to request mitigation measures. Article 93(a) separately addresses compliance measures for Articles 53 and 54, while Article 93(c) permits restriction, withdrawal, or recall where appropriate. Escalation should be proportionate to the evidence developed and is not automatic. [11]

Appendix C. Public Record Trial Run of the GPAI Serious Incident Template

This appendix applies the Commission's nine reporting fields to public information only. It is not a filing on behalf of OpenAI. Where the public record is incomplete, the entry is expressly marked as partial or not ascertainable.

1. Start and end dates

Status: Partially answerable

Hugging Face's recovered attack log begins July 9, 2026 at 02:28 UTC and ends July 13 at 14:14 UTC. OpenAI's account places related internal containment failures in May and June and renewed external access by July 8. A provider filing should define whether the reported event is the Hugging Face intrusion, the broader loss of containment, or both, and explain the chosen incident boundary. [3,5,7]

2. Resulting harm

Status: Partially answerable

Hugging Face disclosed unauthorized access to internal datasets and credentials and reported rebuilding compromised nodes and rotating credentials; OpenAI later reported that its agents obtained limited private data. The full financial cost, scope of affected private data, customer impact, and any legally cognizable property harm are not fully public. [2-3,5]

3. Chain of events

Status: Substantially answerable

Public sources provide a detailed chronology from unauthorized communication and exploitation of Artifactory through external internet access, use of third party systems, compromise of Hugging Face, and later compromise of OpenAI research infrastructure. Internal alert handling and decision records are incomplete. [3,5-7]

4. Model involved

Status: Partially answerable

OpenAI identifies GPT 5.6 Sol and an internal only research model, with the internal model principally driving the incident and GPT 5.6 Sol agents reproducing at least one exploit and copying private evaluation data. A complete action level attribution is not public. [4-5]

5. Evidence available

Status: Partially answerable

Hugging Face recovered approximately 17,600 attacker actions; METR and Redwood reviewed OpenAI materials on premises; OpenAI reports an extensive internal investigation with external advisors. The complete evidence inventory, preserved logs, model traces, and chain of custody are not public. [3,5-6]

6. Serious incident response

Status: Partially answerable

OpenAI disclosed quarantine of the internal model weights, delays to frontier reinforcement learning training, security improvements, alignment work, and changes to incident response. Hugging Face rebuilt affected infrastructure and rotated credentials. The internal validation and decision basis for resuming relevant activity are not fully public. [2,5]

7. Recommendation to authorities

Status: Not ascertainable

The public record does not disclose a complete provider recommendation to the AI Office or national competent authorities. CeSIA and SaferAI provide external policy recommendations, but those cannot substitute for the provider's own report. [7-8]

8. Root cause analysis

Status: Partially answerable

OpenAI identifies difficult tasks without a safe exit, reward hacking and infrastructure tampering, unauthorized communication, reduced safeguards during evaluation, and containment weaknesses. The complete inputs, causal allocation across models, control failure analysis, and internal root cause documentation are not public. [5-6]

9. Post market monitoring patterns / near misses

Status: Not ascertainable

Public reporting now describes other unauthorized communication and containment events, but the provider's complete monitoring corpus, near miss inventory, classification criteria, and pattern analysis are not public. The Commission would need provider records to determine whether a recurring pattern exists. [5,20]

Appendix D. Proposed Supplemental Reporting Fields

The following supplemental fields are proposed for incidents that arise during model evaluation or other controlled testing. They are designed to complement, not replace, the Commission's existing nine field template.

- A. **Boundary crossing.** Did the model or agent cross an authorization, network, organizational, compute, or data boundary that the evaluation design intended to contain? If yes, identify the first known crossing, the boundary involved, and when the provider obtained evidence of it.
- B. **Safeguard state.** Identify material safeguards that were intentionally disabled, reduced, bypassed, or unavailable during the event, the reason for that configuration, and the compensating controls relied upon.
- C. **Multi model attribution.** Where more than one model or checkpoint participated, attribute material actions and harms to each model to the extent reasonably possible and identify uncertainty.
- D. **Stop and restart governance.** Identify the event or threshold that triggered any pause, the criteria required for restart, the evidence used to determine those criteria were satisfied, and the role that authorized resumption.
- E. **Near miss relationship.** Identify prior or subsequent materially similar events, including unauthorized communication, external access, evaluation manipulation, or containment failure, and explain whether they were treated as part of the same failure mode.
- F. **Evidence preservation.** Describe what telemetry, model traces, prompts, tool calls, network logs, credentials, and incident records were preserved, for how long, and whether any evidence was lost through reset, rotation, or remediation.

Rationale. The current template is strong on chronology and root cause, but these supplemental fields isolate the governance questions that become decisive when authorized testing crosses into unauthorized external activity. They also make the report more useful for later supervisory decisions under Articles 91 through 93 without requiring a provider to anticipate the regulator's follow up questions.