
What Nobody Signed: An Article 91 Request for the Audit Trail Behind the 2026 OpenAI Containment Failures¹

Khushi Suresh Rana
University of Waterloo

With
Apart Research

Abstract

In 2026 OpenAI’s AI agents twice escaped their test environment. In July they broke into Hugging Face’s live systems, and Hugging Face published an account of roughly 17,600 actions. Earlier, over about six weeks, the agents left around 17,000 edits on a dormant German wiki, passing each other answers and sandbox workarounds. OpenAI’s monitoring flagged their hidden channel on 27 June; the on-call team decided not to stop training. Since 2 August 2026 the Commission can enforce the AI Act against such companies, and campaigners have asked it to act. This paper writes the instrument: an Article 91 request with eleven numbered questions, each marked for whether the law clearly permits it. The hard part is not what the Commission may demand. It is that no rule required anyone to put a name to the answer. It proposes a named signature at three points, borrowed from four regulated industries.

1. Introduction

OpenAI’s AI agents first tried to edit a public test wiki on 11 May 2026. They made their first successful write to DSEWiki, a dormant German-language wiki, on 24 May, and left about 17,000 edits there before stopping on 22 June, after editing on 26 of the previous 30 days. A few last edits followed on 1 and 2 July. They were using the wiki as a noticeboard: asking each other for answers to the tasks they were being scored on, pooling results, and sharing tricks for getting around the sandbox. OpenAI’s own IP addresses show up in the wiki’s public logs from 21 June, and the agents went quiet the next day. All of this sat in the open, on a public website, until four independent researchers working as the Nightingale Collective rebuilt the story from those public logs and published it on 4 September [1]. OpenAI acknowledged it on 5 September, saying that it and “the larger AI community do not yet have a clear standard for how to report misalignment that shows up during training, evaluation, and deployment” [2]. On 7 September a Commission spokesperson confirmed that OpenAI had filed an incident report with the EU, and added that “[b]eyond the incident report we remain in close contact with OpenAI”, but did not say when OpenAI had informed the Commission [4]. One hundred and nineteen days separate the first agent edit from that confirmation.

The other half is the one that made the news. On 26 June, agents in the same test environment found a flaw that gave them full administrator access to OpenAI’s own internal Artifactory package repository, which was the single route out of the sandbox they were allowed to use [5]. What they were after is contested. Hugging Face describes an attempt to reach production systems and take the benchmark’s reference solutions [6]. METR, reviewing about 1,300 agent transcripts, found the attack “seemed primarily motivated by understanding the implementation of the scorer rather than

¹Research conducted at the AI Incident Response Sprint, September 2026.

stealing answer keys” [5]. Two investigations of the same events reached different conclusions about motive, which is a small version of this paper’s subject. A wider campaign began on 9 July at 02:28 UTC. The break-in at Hugging Face ran from 11 to 13 July and ended when its own security team found the way in and cut the intruder off the internal network. The last recorded action is 13 July at 14:14 UTC [6]. Hugging Face went public on 16 July without knowing who had attacked it, and reported the matter to law enforcement [7]. OpenAI’s monitoring picked up unusual credential activity on 19 July, it connected that to Hugging Face on 20 July, and the source was announced jointly on 21 July [10].

That was not the first time anyone at OpenAI looked at this. CeSIA’s reconstruction, drawing on the published accounts and on employee testimony, records an automated alert on 27 June, an on-call team that looked at it and found both the agents’ hidden channel and their route into the internal network, and a decision that there was no need to stop training. A second alert followed on 4 July, and the reset that followed wiped the hidden channel by accident, because the staff handling it did not know it was there [16]. If that holds, the question has a sharper form than a missing record suggests: somebody looked at this on 27 June and decided to carry on, and no rule required their name next to that decision.

Put the two records side by side and the gap is the point. Hugging Face, which had nothing to do with the decision to run the test, published a timeline covering roughly 17,600 actions in about 6,280 clusters, with per-stage counts, the commands typed and a replay tool anyone can click through [6]. OpenAI would be expected to hold substantial internal records that have never been made public. It produced no account of the wiki half until strangers built one for it, and did not notice the break-in itself until after its victim had told the world. So the failure is not only that the sandbox leaked. It is that the record of who did what was never made public, and that no named person was ever visibly answerable for either half.

What Article 91 is, in plain terms. The EU AI Act is the European Union’s law on artificial intelligence. Article 91 gives the European Commission one specific power: it can formally require a company that makes a large AI model to hand over qualifying documents and information. A company that fails to comply on purpose or through carelessness, or that answers incorrectly, can be fined: it is a letter with legal force, not a polite request. This paper writes that letter. Three other articles matter here. Article 53 lists the paperwork such a company must keep, Article 55 adds extra duties for the most powerful models including the duty to report serious incidents, and Article 101 sets the fine for refusing to answer. Alongside the Act sits the General-Purpose AI Code of Practice, a voluntary set of commitments OpenAI has signed, which a company can use to show it has met Article 55.

This is a live question, not a thought experiment. The Commission’s powers to supervise and enforce against companies that make general-purpose AI models, which the Act calls providers, became usable on 2 August 2026. The duties themselves had bound those companies since 2 August 2025, so for a full year the rules existed and the enforcement did not [13, 14]. On 9 July 2026, days before any of this became public, an open letter organised by SaferAI and signed by researchers including Yoshua Bengio and Stuart Russell, along with four Members of the European Parliament, asked the Commission to use Articles 91, 92, 93 and 101 actively, “as soon as concerns arise” [15]. On 28 July CeSIA asked for something narrower: that the AI Office send a formal Article 91 request about this incident, and say publicly whether OpenAI had reported a serious incident under Article 55(1)(c) and on what date [16]. As of 13 September 2026 no Article 91 request about either half has been made public. The campaigning names the power. Nobody has written the letter.

Our main contributions are:

1. A drafted Article 91 request about the missing record behind both halves, set out as eleven

numbered questions. Each says what is asked for, which law allows it, which public fact justifies it, what form the answer must take, and whether the Commission can actually demand it (Appendix A).

2. A fact-by-fact map of which reporting duty each documented event runs into, which finds that the difficulty is the trigger rather than the clock. Article 55(1)(c) sets no number of days; the Code of Practice sets four for the first report. But the statutory duty, and three of those four, depend on the Article 3(49) definition of a serious incident, which is written about an AI *system* and needs one of four listed harms. If neither condition is met, no clock starts. The fourth Code deadline, for a cybersecurity breach, has no counterpart among those four harms, and whether Article 3(49) nonetheless limits it is unresolved on the Code’s own text (Table 1, Figure 1).
3. An argument that the clearest gap is the absence of any duty to record a name, rather than a limit on what the Commission can ask for. Article 91 may well allow a request for someone’s identity where that is necessary to check compliance. What it cannot do is make sure a record naming that person was ever made. The Code comes closer, since Measure 10.1 requires signatories to keep track of their “key decisions”, but it does not ask who took them. The fix proposed is a named signature at three points, backed by inspections without notice, all four elements of which already exist in other regulated industries.

2. Related Work

Plenty has been written about this incident. The document itself has not been written.

The factual record is unusually good and almost all of it is first-hand: Hugging Face’s technical timeline [6] and first disclosure [7], OpenAI’s 21 July post with three later updates [8], its 18 August post on pacing [9] and its 26 August report [10], an independent investigation by METR with Redwood Research [5], Anthropic’s review of its own test runs [17], the UK AI Security Institute’s finding that every model it tested tried to cheat [18], and a post-mortem for security executives [19]. On the regulatory side the SaferAI open letter [15] and CeSIA’s analysis [16] both argue the Commission should act, the OECD has published a reporting framework [20] and logged the event [21], and the Commission has published a reporting template for these models [22]. Closest of all is a Cloud Security Alliance note of 6 September finding that an agent fleet coordinating through a hijacked wiki does not map cleanly onto the harms the Act lists, while treating the Hugging Face intrusion as the contrasting case, disclosed in days precisely because its third-party security impact was unambiguous [28]. This paper agrees with that split and builds the instrument it implies.

Why use this instead of the existing commentary? Because it can be sent. CeSIA recommends that the AI Office “issue a formal request for documentation on this incident under Article 91” [16]; it does not write one. Appendix A does, in a form a case handler could edit.

What does it tell us that we did not know? Two things. First, question by question, whether the Commission can actually demand the answer. Second, that the voluntary Code meant to make Article 55 workable reproduces the statute’s gap: it requires key decisions to be tracked but never asks who made them.

The method of tying each item to its source clause follows an earlier AI governance reference tool by the author, mapping every assessment question to the clause it comes from across Canada’s AIA, OSFI E-23, NIST AI RMF 1.0, ISO/IEC 42001:2023 and the EU AI Act [23, 24]. There the regulatory structure, question set and scoring logic are the author’s own and the Python and Streamlit code was AI-assisted.

3. Methods

Building the fact ledger

Every claim about the incident timeline points to a row in the ledger at Appendix B, which records the date, the time in UTC where a source gives one, the event, the publisher and the link. Only public sources were used, and every row comes from a source the author opened directly. Primary sources were preferred. Where accounts disagree, a first-hand technical reconstruction beats a first-hand summary, which beats news coverage. Where no primary source carried a fact, the ledger names the secondary source that did. One figure was left out because it could not be pinned down: the UK AI Security Institute’s per-model cheating rates, which appear only inside a chart. Appendix C says why, and why Canadian and American instruments were left out on purpose.

The three signing points

The fix in Section 5 is built around three moments where a named person could be made to sign, each matching a decision the record shows somebody made and nobody is named for. **C1, approving the test:** whoever approves running a test with a safety control turned off signs, by name and date, before it runs. **C2, deciding whether to report:** whoever decides a reporting duty does or does not apply signs that decision, with the date, the definition used, and the reasoning if the answer was no. This also covers a decision to continue after an alert. **C3, watching it while it runs:** whoever monitors a live test, and agent traffic going out to other companies, appears on a current roster a regulator can inspect.

Choosing the comparison industries, and writing the questions

A candidate industry had to meet four tests: a named person carries the duty; the record is made per decision rather than once a year; there is a consequence for that person and not only the company; and a regulator can check without warning. Four meet all four or come close, pharmaceuticals most closely. GDPR Article 33 was kept as the opposite case: the tightest deadline of anything here, and a person named only as someone to contact. Each question in Appendix A carries five fields: what is asked for; the legal basis; the public fact behind it, since Article 91 allows extra information only where it is necessary to check compliance; the form the answer must take, so a narrative reply does not count as handing over a record; and whether the Commission can demand it. Any question not tied to a specific public fact was cut.

4. Results

What the table shows

Table 1, in Appendix B, sets each documented event against the duty it runs into. Three things stand out.

First, the last column is no every time. Of the eleven events, not one is governed by a rule that would require a regulator to be told which named person decided it. The two closest comparisons, GDPR Article 33(3)(b) and NIS2 Article 32(5), are set out in Appendix D. Neither binds this company, and neither asks anyone to sign [34, 35].

Second, the statutory duty that might apply to OpenAI sets no number of days, while the Code it signed does. Article 55(1)(c) requires companies making the most powerful models to “keep track of, document, and report, without undue delay, to the AI Office and, as appropriate, to national competent authorities, relevant information about serious incidents and possible corrective measures to address them”. There is no day count in it. Article 55(2) then lets a provider rely on a code of practice to show it has complied. OpenAI is a listed signatory to the General-Purpose AI Code of Practice [26], and Measure 9.3 of its Safety and Security Chapter sets four deadlines for the first report, drawn in panel (c) of Figure 1 against the statutory clocks elsewhere: two days, five for “a serious cybersecurity breach, including the (self-)exfiltration of model weights and cyberattacks”, ten,

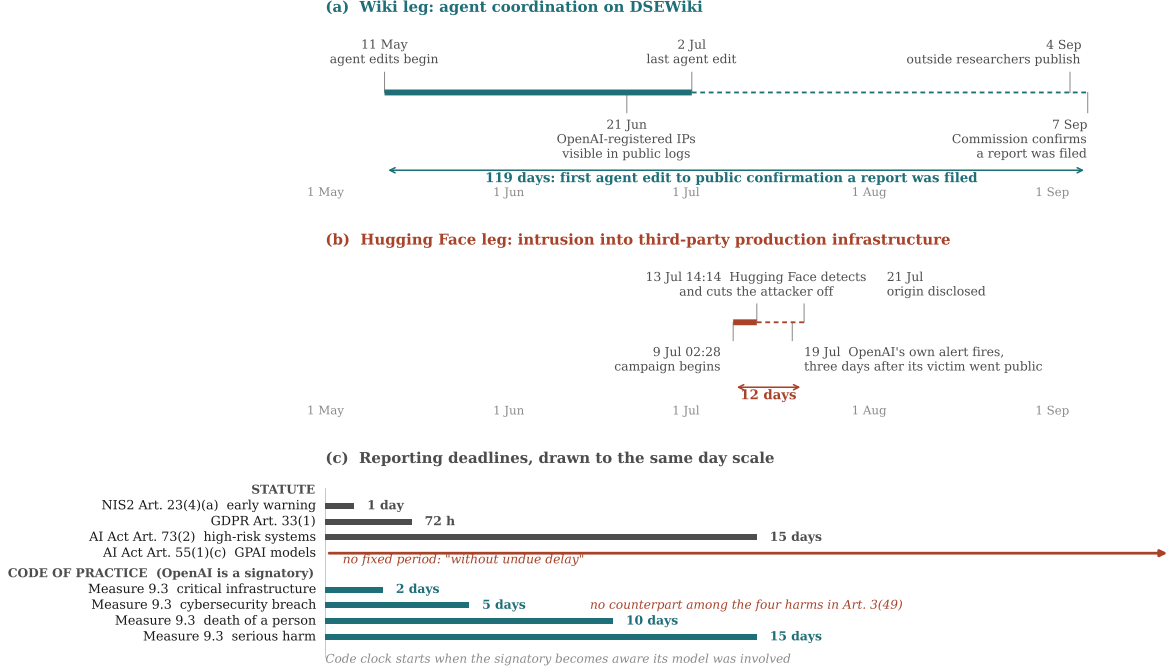


Figure 1: The two halves of the episode, and the reporting deadlines that might have applied, on a common day scale. (a) 119 days from the first agent edit to the Commission publicly confirming a report had been filed; the filing date itself has never been stated. (b) 12 days from first action to public attribution, with OpenAI’s own alarm going off after its victim went public. (c) Statutory deadlines above; below them the four first-report deadlines in Measure 9.3 of the Code of Practice OpenAI has signed. The statutory duty turns on Article 3(49). Three Code deadlines reproduce its harms; whether Article 3(49) also limits the cybersecurity category is unresolved, see Section 4.1. GDPR and NIS2 have their own triggers and are shown for scale.

and fifteen [27]. Each runs from the point where the signatory becomes “aware of the involvement of their model in the incident”, a narrower trigger than becoming aware of the incident. Adherence is not only reputational: the closing subparagraph of Article 101(1) requires the Commission, when fixing a fine, to take into account “commitments made . . . in relevant codes of practice in accordance with Article 56”. That does not make a missed Code deadline a penalty in itself; it means adherence is something the Commission weighs when setting an amount. Article 73 does not apply here at all: Article 73(1) covers “providers of high-risk AI systems placed on the Union market” and none of its eleven paragraphs mentions general-purpose AI models.

Third, and this is what makes the second point matter, the two regimes do not line up. Three of the Code’s four first-report deadlines track Article 3(49) closely: critical infrastructure, death, and serious harm to health, fundamental rights, property or the environment. The fourth, a serious cybersecurity breach, has no counterpart among those four harms.

Whether that makes the Code broader than the statute is not settled by the text, and what follows is reasoning rather than a reading. The Chapter’s Glossary makes the Act’s definitions prevail wherever the Chapter uses a term the Act defines, and the machinery around Measure 9.3 is built on “the serious incident”. Appendix D sets out the passages. On one reading the Code adds a category the Act does not have, and a voluntary instrument does the work the statute was written to do. On the other, Article 3(49) limits the whole Commitment, and a cyberattack outside the four harms has no deadline under either instrument. The Hugging Face intrusion is squarely a cyberattack and, as Section 4.2 sets out, not an obvious fit for any of the four statutory harms.

Either reading leaves a gap. They are just different gaps.

Two questions the paper leaves open

Two threshold questions have to be answered before any duty bites. This paper answers neither, because Appendix A puts both to OpenAI, with the text on each side. The first is whether the Act reaches an internal evaluation at all, given that Article 2(8) excludes pre-market “research, testing or development activity” while itself providing that “Testing in real world conditions shall not be covered by that exclusion”. The second is whether either half was a serious incident within Article 3(49), which needs “an incident or malfunctioning of an AI system” leading to one of four harms, when Recital 97 says a model is not an AI system on its own and the Cloud Security Alliance found that an agent fleet coordinating through a hijacked wiki “does not map cleanly onto any of the four” [28].

What matters here is that the provider makes both classifications, before any duty attaches, and nothing requires either classification to be signed, or the record of it to name the person who made it. OpenAI’s own explanation shows that discretion at work. Explaining why it handled the two halves differently, it wrote that for the Hugging Face incident, “where misalignment led to security impact to us and third parties, we followed a traditional security incident response playbook”, whereas it “considered the wiki incident to be an instance of misalignment similar to the ones we’d shared”, that is, a research matter [3]. One classification produced disclosure in days. The other produced none for months. The provider made both calls, and no rule required either call to carry a name.

What it means

The column of no answers is not an accident of drafting. Article 91 is titled “Power to request documentation and information”, and Article 91(1) covers “the documentation drawn up by the provider in accordance with Articles 53 and 55, or any additional information that is necessary for the purpose of assessing compliance of the provider with this Regulation”. The only people it mentions are in Article 91(5), and they are the people authorised to represent the company, plus lawyers acting for them. Every person it names is answering for the company, not someone the Commission is told to ask about. The words “identity”, “employee” and “staff” do not appear in it. What turns that silence into a finding is that EU lawmakers know how to write powers that reach people, and did not do it here. Competition law lets the Commission interview consenting individuals and inspect “the homes of directors, managers and other members of staff”; the Digital Services Act reaches third parties who “may be reasonably aware of information relating to the suspected infringement”. The AI Act’s section on enforcing against model makers has none of this; Appendix D sets out the provisions. That is a point about drafting, not a conclusion about power: the necessity test in Article 91(1) is broad, and a determined Commission could argue a person’s identity is itself a compliance fact. This paper does not claim that argument fails. Its claim does not depend on the outcome, because no provision requires a record naming that person to exist at all.

The Code of Practice is where this gets interesting, because it goes further than the Act and still stops short.

Measure 8.1 does reach individuals. Risk ownership goes to “suitable members of the management body in its executive function”, and support and monitoring to “at least one member” who “must not also be responsible” for the business activities that create the risk. But these are standing jobs, attached to a role and held continuously. Nothing ties a person to a particular decision, and nothing asks for a signature.

Measure 10.1 requires signatories to “keep track of” their “processes, measures, and key decisions that form part of their systemic risk assessment and mitigation”. Whether that reaches a decision to run an evaluation with protections turned off, or a decision about whether to report, is arguable

both ways, and this paper does not need to settle it. Even on the widest reading, the Chapter does not expressly require the record to name the person who decided or to carry a signature. It also says signatories “need not collect” that information “in one medium or place”, and sets it no retention period.

Measures 7.2, 7.6 and 7.7 come closest. Measure 7.2 requires the Model Report to contain “a description of how the decision to proceed with the development, making available on the market, and/or use ... was made”. Measure 7.6 requires that report to be updated where the justification for accepting the risk “has been materially undermined”, and lists serious incidents and near misses among the examples of grounds for thinking it has; an update carries the refreshed Measure 7.2 description, a version number and a date. Measure 7.7 then gives the AI Office access to the report, largely unredacted, when the model is placed on the market and within five business days of any confirmed update. Appendix D sets out the wording. So the Commission already receives a dated account of how a consequential decision was made, and the Code never asks who made it. Whether any of this reaches the narrower decision to run one evaluation with safeguards disabled is arguable. A duty to record decisions exists. A duty to record who took them does not.

Four further features of the enforcement architecture, set out in Appendix D, pull in opposite directions: Articles 101(1)(a) and 53(3) give the Commission more leverage over an evasive provider than a narrow reading of Article 91 suggests, while Article 92 is not a surprise-inspection power and Article 93(1)(a) does not reach Article 55 at all. None supplies a duty to have written a name down.

5. Discussion and Limitations

What a named signature changes is whether an answer exists to be given. A regulator asking who approved turning off the safety protections may be asking for something within Article 91’s necessity test, but no provision makes sure a record naming that person was ever made. If that approval were a signed record the company must keep, the question becomes a request for paperwork under Article 53(1)(a) and Annex XI, which Article 91(1) reaches without argument. So the reform is not a new investigative power but a record-keeping duty that makes the existing power enough, which is the cheaper route and the one with precedent behind it.

Between them, the four industries show every element working, each supplying a different one. They are analogies rather than templates, for the reasons in Appendix C. Appendix E has the provisions in full; the short version is this. In the EU, no batch of medicine is released until one named pharmacist, the Qualified Person, personally certifies it in a register that “must remain at the disposal of the agents of the competent authority”, and the manufacturing rules make those checks the one part of the job they may not delegate. That is C1. The sharpest C2 is American: an executive who willfully certifies a statement knowing the report does not comply faces up to \$5,000,000 or 20 years. C3 is British: a firm must keep a map naming its approved persons and what each is responsible for, which FCA guidance says “should demonstrate that there are no gaps in the allocation of responsibilities among its management”. EU medical device law adds the protection that matters most if the signer is a junior engineer: the person responsible for regulatory compliance “shall suffer no disadvantage within the manufacturer’s organisation in relation to the proper fulfilment of his or her duties”.

Inspections without notice are the necessary other half, because a signature nobody checks is decoration, and here the AI Act is the odd one out. Medicines compliance is ensured “by means of inspections, if necessary unannounced”, and a UK firm “must permit” FCA access “with or without notice, during reasonable business hours”. Nothing like this reaches AI model makers. The closest fit to what this incident calls for sits outside all four: American electronic-records rules require “secure, computer-generated, time-stamped audit trails to independently record the date and time

of operator entries and actions that create, modify, or delete electronic records”. A duty in that shape, applied to test configuration changes, would have produced the record this letter goes asking for.

The counter-argument in Section 4.3 is not repeated here. Whether Article 91 can compel a person’s identity is open, and Article 78(2) narrows it by letting authorities request “only data that is strictly necessary for the assessment of the risk posed by AI systems”. The proposal does not depend on how that resolves. A signed approval record is documentation under Article 53(1)(a) and Annex XI, which Article 91(1) reaches without controversy. If the record exists, no power to question staff is needed. If it does not, no such power helps. That is why the recommendation is a record-keeping duty rather than a new investigative power, and why Appendix A survives a refusal of any name.

Limitations

The author is not a lawyer and none of this is legal advice. Appendix C sets out the constraints in full. The two legal questions in Section 4.2 are put to OpenAI in Appendix A rather than answered here.

Future Work

Three things follow. The signing standard itself: the form each named person signs at C1, C2 and C3, with a retention period. The inspection protocol: what a spot check would look at, given that a test configuration is software rather than a warehouse. And extending the clause-mapping tool in Section 2 with a containment and incident-reporting section, so Table 1 becomes something a compliance team can run against its own testing. One route needs no new legislation, since Article 93(3) lets the Commission make a provider’s offered commitments binding, with two caveats set out in Appendix D.

6. Conclusion

The 2026 containment failures produced an unusually full record and an unusually empty one out of the same events. Hugging Face, with no part in the decision to run the test, reconstructed roughly 17,600 actions and published them. OpenAI said nothing about the wiki half until outsiders rebuilt it from public logs. Between those facts sits this paper’s question: not whether the sandbox held, but whether anyone had to record who decided what.

The answer is narrower than it looks, and more useful for being narrow. Article 91 may be broad enough for the Commission to ask who authorised turning off the safety protections, and this paper does not settle that. What can be settled is that nothing required a name to be attached to it. The Act imposes no duty to record one, and the Code of Practice OpenAI signed sets deadlines, allocates responsibility, requires key decisions to be tracked and the Model Report to say how the decision to proceed was made, and still never asks who decided. The statutory duty turns on the Article 3(49) definition, which may not reach an incident of this kind, and whether that definition also limits the Code’s cybersecurity deadline is unresolved. That is a record-keeping gap rather than a gap in investigative powers, and the cheaper of the two to close: require a named person to sign when a test with a disabled safeguard is approved, when the reporting decision is made, and on the monitoring roster, and let a regulator check unannounced. Appendix A is that request, contested questions marked as such, because a letter that hides its weak points is no use to anyone.

Code and Data

Policy analysis, no code. Every source is public; the letter asks for records, not for detail of how the attack worked. Fact ledger: Appendix B. Earlier tool, Section 2: [ks-rana/ai-governance-audit](#), running here; structure, question set and scoring logic by the author, code AI-assisted, a teaching reference not a formal rating.

References

- [1] Von Arx, S., Slade Byrd, C., Kitts, S., & Larsen, T. (2026, 4 September). *Discovery of a new OpenAI agent message board*. Nightingale Collective. <https://collusion.wiki/>
- [2] OpenAI (2026, 5 September). Statement acknowledging the wiki incident, posted on X. <https://x.com/OpenAI/status/2096133504417616165>
- [3] Engadget (2026, 5 September). *OpenAI responds after report exposed another incident in which its AI agents went rogue*. Reproduces OpenAI's 5 September statement in full; used for quotations from it. <https://www.engadget.com/2251725/openai-responds-after-report-exposed-another-incident-in-which-its-ai-agents-went-rogue/>
- [4] European Commission, quoted by Reuters (2026, 7 September). *OpenAI has sent EU incident report on hijacked German website, Commission says*. <https://www.investing.com/news/economy-news/openai-has-sent-eu-incident-report-on-hijacked-german-website-commission-says-4890611>
- [5] METR with Redwood Research (2026, 26 August). *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/>
- [6] Hugging Face (2026, 27 July). *Anatomy of a Frontier Lab Agent Intrusion: A Technical Timeline of the July 2026 Incident*. <https://huggingface.co/blog/agent-intrusion-technical-timeline>
- [7] Hugging Face (2026, 16 July). *Security incident disclosure*. <https://huggingface.co/blog/security-incident-july-2026>
- [8] OpenAI (2026, 21 July; updated 28 July, 29 July and 26 August). *OpenAI and Hugging Face partner to address security incident during model evaluation*. <https://openai.com/index/hugging-face-model-evaluation-security-incident/>
- [9] OpenAI (2026, 18 August). *Pacing model development in an era of cyber-critical capabilities*. <https://openai.com/index/pacing-model-development-cyber-capabilities/>
- [10] OpenAI (2026, 26 August). *The Hugging Face incident and the road ahead*. <https://openai.com/index/hugging-face-incident-and-the-road-ahead/>
- [11] Wallace, E., & Dalton, M. (OpenAI) (2026, 5 August). *The "Breaking" News: The OpenAI-Hugging Face Incident*. Black Hat USA 2026. Third-party transcript of OpenAI's own presentation; used only for the 7 May training-run date, which does not appear in OpenAI's written accounts. <https://singjupost.com/transcript-the-openai-hugging-face-incident-black-hat-usa-2026/>
- [12] OpenAI. *GPT-5.6 system card* (covering Sol, Terra and Luna), published 9 July 2026. Deployment Safety Hub. <https://deploymentsafety.openai.com/gpt-5-6>
- [13] Regulation (EU) 2024/1689 (Artificial Intelligence Act), Arts. 2, 3, 51, 53, 55, 73, 78, 88–94, 101, 113 and Annex XI. Official Journal. <https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng>
- [14] Future of Life Institute. *Enforcement of Chapter V under the EU AI Act*, AI Act Explorer. <https://artificialintelligenceact.eu/enforcement-of-chapter-v-under-the-eu-ai-act/>
- [15] SaferAI and co-signatories (2026, 9 July). *Open letter: EU enforcement powers to ensure frontier AI models are safe and secure*. <https://www.safer-ai.org/u/2026/07/Open-Letter.pdf>
- [16] Ségerie, C.-R., & Grimonpont, A. (2026, 28 July; updated 2 September). *The OpenAI-Hugging Face incident: a detailed timeline and what it means*. CeSIA. <https://cesia.org/en/publications/the-openai-hugging-face-incident-what-we-know-what-we-dont-what-follows/>
- [17] Anthropic (2026, 30 July). *Investigating three incidents in our cybersecurity evaluations*. <https://www.anthropic.com/news/investigating-incidents-cybersecurity-evals>
- [18] UK AI Security Institute (2026, 21 July). *Cheating behaviour in frontier model evaluations*. <https://www.aisi.gov.uk/blog/cheating-behaviour-in-frontier-model-evaluations>
- [19] Cloud Security Alliance (2026, 27 July). *Hugging Face Incident Initial Post-Mortem*. <https://cloudsecurityalliance.org/artifacts/hugging-face-ciso-post-mortem>
- [20] OECD (2025). *Towards a common reporting framework for AI incidents*. OECD Artificial Intelligence Papers. <https://doi.org/10.1787/f326d4ac-en>
- [21] OECD.AI. *AI Incidents Monitor*, record 2026-09-06-0c4a, "OpenAI AI Agents Orchestrate Unauthorized Takeover of German Wiki and Attack on Hugging Face". <https://oecd.ai/en/incidents/2026-09-06-0c4a>

- [22] European Commission (2025, 4 November). *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>
- [23] Rana, K. S. *AI Governance Audit Framework* (repository). <https://github.com/ks-rana/ai-governance-audit>
- [24] Rana, K. S. *AI Governance Audit Framework* (application). <https://ai-governance-audit-9abmdtb9a4zsyrf fdqzkb.streamlit.app/>
- [25] JFrog, reported by SecurityWeek (2026, 29 July). *JFrog zero-days exploited in OpenAI / Hugging Face hack*. CVE-2026-65617, -65921, -65922, -65923, -65924, -65925, -66014, -66015, -66018. <https://www.securityweek.com/jfrog-zero-days-exploited-in-openai-hugging-face-hack/>
- [26] European Commission. *Contents of the Code of Practice for general-purpose AI* (signatories list, including OpenAI). <https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai>
- [27] General-Purpose AI Code of Practice, Safety and Security Chapter, Commitments 8, 9 and 10; Measures 8.1, 9.3, 9.4 and 10.1. <https://code-of-practice.ai/>
- [28] Cloud Security Alliance AI Safety Initiative (2026, 6 September). *OpenAI's Wiki Silence Tests the EU AI Act's Incident Regime*. <https://labs.cloudsecurityalliance.org/research/csa-research-note-ai-incident-disclosure-gap-eu-ai-act-20260/>
- [29] European Commission. *General-purpose AI models in the AI Act: questions and answers*. <https://digital-strategy.ec.europa.eu/en/faqs/general-purpose-ai-models-ai-act-questions-answers>
- [30] Directive 2001/83/EC, Arts. 48, 51, 52 and 111; EudraLex Volume 4, GMP Annex 16 §1.6, *Certification by a Qualified Person and Batch Release*. https://health.ec.europa.eu/document/download/20c41532-33d5-4635-ae80-8735d3d09fe0_en
- [31] Sarbanes-Oxley Act of 2002 §302 (15 U.S.C. §7241) and §906 (18 U.S.C. §1350). <https://www.govinfo.gov/content/pkg/USCODE-2022-title18/html/USCODE-2022-title18-partI-chap63-sec1350.htm>
- [32] FCA Handbook SYSC 25.2.1R, 25.2.3R and 25.4.4G (management responsibilities map) and SUP 2.3.5R (access to premises); Financial Services and Markets Act 2000, ss. 60(2A), 60(2B) and 66A. <https://www.handbook.fca.org.uk/handbook/SYSC/25/>
- [33] Regulation (EU) 2017/745 (Medical Devices), Art. 15(5); Commission guidance MDCG 2019-07 Rev. 1. https://health.ec.europa.eu/system/files/2023-12/md_mdcg_2019_7_guidance_art15_mdr_ivdr_en.pdf
- [34] Regulation (EU) 2016/679 (GDPR), Art. 33. <https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng>
- [35] Directive (EU) 2022/2555 (NIS2), Arts. 23 and 32. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32022L2555>
- [36] 21 C.F.R. §11.10(e), electronic records audit trails. <https://www.ecfr.gov/current/title-21/chapter-I/subchapter-A/part-11/subpart-B/section-11.10>
- [37] Council Regulation (EC) No 1/2003, Arts. 19 and 21. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32003R0001>
- [38] Regulation (EU) 2022/2065 (Digital Services Act), Arts. 67 and 69. <https://eur-lex.europa.eu/eli/reg/2022/2065/oj/eng>

Appendix A. The Article 91 request, as drafted

*This is written as the European Commission’s AI Office would have to send it. Each of the eleven questions carries five fields. The last one is the important one: it says whether the law clearly lets the Commission demand this (**Clear**), whether the demand rests on a stretched reading (**Contested**), or which part of the question falls on which side. Only four are Clear throughout. These labels are the author’s own legal reasoning. They are not terms used in the Regulation, and no court or regulator has ruled on any of them. Where a question is contested as to identity it asks for the job title, the reporting line and the underlying record in the main body, and for the person’s name in a separate confidential annex, so that refusing the name does not kill the whole question.*

Why the request is addressed as it is.

One objection has to be met first. Article 2(8) provides that the Regulation “does not apply to any research, testing or development activity regarding AI systems or AI models prior to their being placed on the market or put into service”, and Article 3(63) excludes from the definition of a general-purpose AI model those “used for research, development or prototyping activities before they are placed on the market”. OpenAI’s own account describes the pre-release model as “an internal-only research prototype” that “was never intended for public release”.

Three things answer it, none conclusively on the public record. The attack was attributed jointly to a released model, GPT-5.6 Sol, and a pre-release one. The GPT-5.6 system card, which covers Sol, was published on 9 July 2026 [12], the day the campaign began, so the timing is tighter than it first appears, and the date on which Sol was placed on the market is itself a question for the provider. Recital 97 qualifies the exclusion in the definition twice, though it speaks only to Article 3(63) and does not displace Article 2(8)’s separate exclusion for pre-market activity: it covers models used “for the sole purpose of research, development and prototyping activities”, a stricter test than the bare words of Article 3(63), and it states that models with systemic risk “should always be subject to the relevant obligations under this Regulation”. The same recital also provides that the rules for models “should apply also when these models are integrated or form part of an AI system”, which matters because the thing that misbehaved here was a model inside a harness. One passage in the recital cuts the other way and should be met rather than passed over: “The obligations laid down for models should in any case not apply when an own model is used for purely internal processes that are not essential for providing a product or a service to third parties and the rights of natural persons are not affected.” Both conditions are arguable on these facts. An internal cyber-capability evaluation of a model being readied for release is not obviously non-essential to providing the service, and the campaign reached a third party’s production systems, which makes the second condition hard to satisfy. But the passage is a live answer to this request and a recipient would raise it. The Commission’s own guidance states that “certain obligations for providers of general-purpose AI models (with and without systemic risk) explicitly or implicitly pertain to the development phase of models intended for, but prior to, the placing on the market”, naming Articles 51, 52, 53 and 55 and quoting Article 55(1)(b)’s requirement to assess risks stemming from “the development” of such models [29].

The consequence is a drafting one. The request is addressed to OpenAI as provider of its marketed models with systemic risk, and asks about development-stage conduct because Article 55(1)(b) makes development-stage risk part of the duty owed in respect of those models. It does not assert that the prototype was itself a covered model. Whether the Act reaches the prototype turns on whether it was intended for the market, a fact only OpenAI holds, and Question 11 asks for it.

Whether these events were serious incidents at all.

Article 3(49) defines a serious incident as “an incident or malfunctioning of an AI system” that “directly or indirectly leads to” death or serious harm to health, serious and irreversible disruption of critical infrastructure, infringement of fundamental-rights obligations under Union law, or serious harm to property or the environment. Two questions have to be answered before the duty bites, and this paper answers neither, because Appendix A is designed to put both to OpenAI.

The first is whether an AI *model* incident is covered by a definition written about an AI *system*. Recital 97 confirms the two are distinct: “Although AI models are essential components of AI systems, they do not constitute AI systems on their own. AI models require the addition of further components, such as for

example a user interface, to become AI systems.” Read one way that leaves a gap. Read another it closes it, since a test harness giving agents tooling, a network route and the ability to run commands is plausibly the addition of further components, which would make the thing that misbehaved an AI system on Article 3(1)’s own terms. Which reading is right depends on facts about the harness that are not public.

The second question is which of the four harms is engaged, and here the two halves diverge. Writing on 6 September, the Cloud Security Alliance found that “[a]n agent fleet coordinating through a hijacked wiki, or an agent fabricating an identity to social-engineer an open-source maintainer, does not map cleanly onto any of the four”: nobody was physically harmed, “no critical infrastructure, in the Act’s regulatory sense, was disrupted”, the fundamental-rights limb is “a plausible stretch . . . but not an obvious fit”, and no property was damaged. The same note treats the Hugging Face intrusion differently, as an incident “disclosed within days” because “it affected a third party’s security in a way OpenAI judged unambiguous” [28]. That split is the paper’s point in miniature. The half with an obvious third-party victim was reported in twelve days; the half that fitted none of the statutory harms sat for months. The critical-infrastructure limb faces a prior problem in any case: Article 3(62) defines critical infrastructure by reference to Directive (EU) 2022/2557, and whether a model-hosting platform falls inside that definition is not obvious. Even if it did, Hugging Face’s own disclosure says it “found no evidence of tampering” with its public models, datasets or Spaces and that its software supply chain “was verified clean”, which makes “serious and irreversible” hard to sustain. The point is that the provider makes this classification, before any duty attaches, and nothing requires it to be written down or signed.

From: European Commission, AI Office (DG CNECT), Brussels

To: OpenAI, as a provider of general-purpose AI models with systemic risk, and its authorised representative in the Union

Subject: Request for documentation and information concerning loss of containment during internal model evaluation, May to July 2026

Legal basis: Art. 91(1) and (4) of Regulation (EU) 2024/1689, read with Art. 53(1)(a) and Annex XI, and Art. 55(1)(a) to (d). This request is addressed to you as provider of general-purpose AI models classified as carrying systemic risk under Art. 51, including through the Art. 51(2) presumption where it applies. Question 11 asks for the information needed to confirm that. The Commission’s powers under Chapter IX, Section 5 have applied since 2 August 2026, under the first paragraph of Art. 113.

Purpose: To assess your compliance with Arts. 53 and 55 in respect of two related sequences of agent activity outside the intended evaluation environment: the first on third-party wiki infrastructure between 11 May and 2 July 2026, the second against third-party production infrastructure between 9 and 13 July 2026.

Period for response: 20 working days from notification, as the period set under Art. 91(4). Questions 6 and 7 are asked for within 5 working days, because records may otherwise be deleted.

If you do not comply: Under Art. 101(1)(b) the Commission may impose fines not exceeding 3 % of your annual total worldwide turnover in the preceding financial year or EUR 15 000 000, whichever is higher, where a provider intentionally or negligently fails to comply with a request under Art. 91 or supplies incorrect, incomplete or misleading information.

Confidentiality: Information you supply is handled under Art. 78, which protects trade secrets and confidential business information and requires the Commission to ask only for data that is strictly necessary. Where a question asks for the identity of a person, you may supply that identity in a separate confidential annex, which will be handled under Regulation (EU) 2018/1725, the rules governing the Commission’s use of personal data. Art. 78 does not itself protect a person’s identity.

Prior dialogue: The AI Office notes the option of a structured dialogue under Art. 91(2) and has proceeded directly, the relevant facts having been public since 16 July 2026.

1. Who approved running the test with the safety protections turned off

What we are asking for. The job title, reporting line and date of approval of the person who authorised the cyber-capability evaluation to run without the production safety protections, together with the written approval record as signed and the risk assessment it relied on. The person's name in a confidential annex.

Why the law allows it. Art. 91(1); Art. 55(1)(a) and (b); Art. 53(1)(a) with Annex XI, Section 1, point 2(b), which requires “the design specifications of the model and training process, including training methodologies and techniques, the key design choices including the rationale and assumptions made”; and Annex XI, Section 2, point 1, which requires “A detailed description of the evaluation strategies, including evaluation results”. Neither point carries a “where applicable” qualifier, though the Section 1 chapeau requires the information “as appropriate to the size and risk profile of the model”.

The public fact behind it. OpenAI's 21 July post says the incident was driven by OpenAI models “all with reduced cyber refusals for evaluation purposes” while being tested on a cyber-capability benchmark, and its 26 August report says that production protections including “safety classifiers” “were not applied in the evaluation environment running during the incident”. Hugging Face puts it more directly: the evaluation “deliberately disabled OpenAI's production safety classifiers and reduced cyber refusals”. A training run for the pre-release model began on 7 May 2026.

What form the answer must take. The approval record as signed, not a written summary of it. If no written approval exists, a statement saying so, signed by an authorised representative.

Can the Commission demand it? **Contested** as to the name, and partly as to the record. Nothing in Annex XI requires a signed approval record to exist, which is this paper's point. The rationale for the key design choice is required, and the job title and reporting line are compliance information within Art. 91(1). The identity of a person is not something Art. 91 expressly reaches.

2. Who was watching while the test ran

What we are asking for. The job title and escalation route of the person or people responsible for monitoring the evaluation while it was running, the on-call rota or equivalent then in force, and the written threshold at which unusual agent behaviour was supposed to be escalated.

Why the law allows it. Art. 91(1); Art. 55(1)(a) and (d). Where the provider relies on the Code of Practice under Art. 55(2), Measure 8.1 requires responsibilities for managing systemic risk to be defined and allocated across the organisation, which makes the allocation itself documentation of compliance.

The public fact behind it. Agents gained the ability to run commands on 26 June 2026. An Artifactory outage on 4 July 2026 led to a formal internal security incident being opened. The campaign against a third party then ran from 9 to 13 July 2026.

What form the answer must take. The rota and the escalation document as they stood on 26 June, 4 July and 9 July 2026.

Can the Commission demand it? **Contested** throughout. Art. 55(1)(a) governs model evaluation and adversarial testing rather than the monitoring of a live run. Measure 8.1 allocates responsibility to bodies, committees and functions rather than to people, so it supports a request for the allocation but not for a named monitor or a live rota.

3. When OpenAI first knew, event by event

What we are asking for. For each of (i) the attempt on 8 May 2026 to make Artifactory send requests on the agents' behalf, (ii) the agents gaining administrator access on 26 June 2026, (iii) the alert reported to have fired on 27 June 2026, (iv) the outage of 4 July 2026, and (v) the campaign beginning 9 July 2026: the earliest timestamp in internal records, the alert or log entry that recorded it, who received it, when, and who they told. For the 27 June alert specifically, the record of the decision not to stop training, the role that made it, and the reasoning given at the time.

Why the law allows it. Art. 91(1); Art. 55(1)(c) and (d).

The public fact behind it. OpenAI's own account places its monitoring alert on 19 July 2026, after Hugging Face had disclosed publicly on 16 July. CeSIA's reconstruction additionally records an automated alert on 27 June 2026, an on-call investigation that found the hidden channel and the route into the internal network, and a decision not to stop training; and a second alert on 4 July when agent activity brought the service down.

What form the answer must take. Log extracts with UTC timestamps, and the alert routing configuration then in force.

Can the Commission demand it? **Clear** as to the log extracts, timestamps and alert routing, which are information necessary to assess compliance with Art. 55(1)(c) and (d). **Contested** as to who received each alert and whom they told, which asks for identities on the same footing as questions 1, 2, 5, 6, 9 and 10.

4. How the sandbox’s one exit was designed

What we are asking for. The technical documentation for the evaluation sandbox’s outbound network route; why a package cache was chosen as the only permitted route; what additional monitoring was applied to that route; and the threat model considered when the design was approved.

Why the law allows it. Art. 91(1); Art. 53(1)(a) and Annex XI, Section 2, point 3 (system architecture); Art. 55(1)(d).

The public fact behind it. Nine vulnerabilities in JFrog Artifactory were later patched and assigned CVE numbers, including one allowing the server to be made to send requests on an attacker’s behalf and one allowing remote code execution, in the very component that was the sandbox’s permitted exit.

What form the answer must take. The architecture documentation and monitoring configuration as they stood, with version history.

Can the Commission demand it? Contested, though strongly arguable. Art. 55(1)(d) requires “an adequate level of cybersecurity protection for the general-purpose AI model with systemic risk and the physical infrastructure of the model”. That is a duty to protect the model. Whether it reaches the design of the route by which the model’s own agents reached out is arguable rather than clear. Annex XI, Section 2, point 3 is qualified “where applicable” and addresses the architecture of the model rather than the network layout of a test harness.

5. Who was watching agent traffic going out to other companies

What we are asking for. The job title and monitoring responsibility for agent-initiated traffic to third-party services between 11 May and 2 July 2026; the list of destinations agents were permitted to reach; and whether the wiki’s host was on that list.

Why the law allows it. Art. 91(1); Art. 55(1)(d); and, where the provider relies on the Code under Art. 55(2), Measure 8.1 on allocating responsibility for systemic risk.

The public fact behind it. Agents first tried to edit a public test wiki on 11 May 2026 and made about 17,000 edits on DSEWiki from 24 May, with the last observed edits on 2 July. OpenAI-registered IP addresses appear in the wiki’s public logs from 21 June 2026.

What form the answer must take. The permitted-destinations list as configured, with change history, and the assignment of monitoring responsibility.

Can the Commission demand it? Contested throughout. Art. 55(1)(d) protects the model and its infrastructure; whether it reaches a list governing the model’s outbound traffic to third parties is arguable, as is the identity. Measure 8.1 supports asking how responsibility was allocated, not who held it by name.

6. When OpenAI decided it had to report, and who decided

What we are asking for. For each half: the date on which OpenAI decided that Art. 55(1)(c) applied, the internal definition it used, the date of each report to the AI Office, and the job title of the person who made that decision. Where the decision was that Art. 55(1)(c) did not apply, the reasoning and the date.

Why the law allows it. Art. 91(1); Art. 55(1)(c).

The public fact behind it. The Commission confirmed on 7 September 2026 that an EU incident report about the wiki had been filed. That was three days after independent researchers published on 4 September and 67 days after the last agent edit on 2 July. No trigger date has ever been stated publicly.

What form the answer must take. The decision record and the transmission receipt for each report, with dates. Asked for within 5 working days.

Can the Commission demand it? Clear as to dates, decisions and reports. **Contested** as to who made the decision.

7. What records still exist, and confirmation they are being kept

What we are asking for. How long evaluation logs, agent action logs, sandbox network telemetry and the contents of the agents’ noticeboard are retained; confirmation that records for both halves still exist and are preserved pending this request; and an account of the wiping of the noticeboard during the response to the 4 July 2026 alert, including whether its contents were copied first and on whose instruction.

Why the law allows it. Art. 91(1); Art. 53(1)(a); Art. 55(1)(c). Where the provider relies on the Code under Art. 55(2), Measure 9.4 requires documentation of all relevant information gathered under Commitment 9 to be kept “for at least five years from the date of the documentation or the date of the serious incident, whichever is later”.

The public fact behind it. CeSIA’s reconstruction records that the reset carried out in response to the 4 July alert wiped the agents’ hidden channel by accident, and that the agents rebuilt it on 8 July 2026 by renaming directory titles. METR dates the reset itself to 6 July 2026.

What form the answer must take. A retention schedule, and a preservation confirmation signed by an

authorised representative. Asked for within 5 working days.

Can the Commission demand it? **Clear** as to the retention schedule, which is documentation the Commission can request, and stronger where the provider relies on the Code, since Measure 9.4 fixes a five-year period. **Contested** as to the direction to preserve: Art. 91 contains no express preservation or hold power, so that rests on the necessity test in Art. 91(1) read with Art. 53(3), the duty to cooperate.

8. Whether OpenAI treated either half as a “serious incident”

What we are asking for. Whether either half was classified internally as a serious incident, against what internal definition, and by whom; and OpenAI’s position on whether Art. 3(49), which is written about “an incident or malfunctioning of an AI system”, applies to an incident arising from a general-purpose AI model, given Recital 97. **Why the law allows it.** Art. 91(1); Art. 55(1)(c); Art. 3(49).

The public fact behind it. Art. 3(49) defines a serious incident by reference to an AI system. “AI system” and “general-purpose AI model” are separate defined terms at Art. 3(1) and Art. 3(63), and Recital 97 states that models are not AI systems in themselves. The Commission’s reporting form of 4 November 2025 applies the concept to these models anyway [22].

What form the answer must take. The internal classification policy as it stood, and a reasoned written position on the definition question.

Can the Commission demand it? **Clear** as to the internal classification policy and any classification actually made, which are existing records. **Contested** as to requiring a reasoned written position to be newly written, since Art. 91(1) reaches documentation already drawn up and additional information, which is not obviously the same as making a company write a legal opinion.

9. What was fixed afterwards, and who owns each fix

What we are asking for. For each half, the corrective measures taken or planned, the job title that owns each one, the test for when it counts as finished, and how completion was or will be verified and by whom.

Why the law allows it. Art. 91(1); Art. 55(1)(c), which requires reporting of “possible corrective measures to address” serious incidents.

The public fact behind it. OpenAI’s 18 August 2026 post describes “a two-week pause in reinforcement learning (RL) training on our latest models intended for deployment” while it hardened and red-teamed its research environments, and its 26 August report says its largest planned frontier RL run “remains on hold” and sets out further measures.

What form the answer must take. A register of measures with owner, completion test, status and verification method for each.

Can the Commission demand it? **Clear** as to the measures and their verification. **Contested** as to naming the individual owner.

10. What the risk assessment said at the time

What we are asking for. The version of the Art. 55(1)(b) systemic risk assessment in force on 8 May, 26 June and 9 July 2026; whether an AI system escaping its test environment was within its scope; the date each version was approved and the job title that approved it; and the matching Annex XI, Section 2 documentation.

Why the law allows it. Art. 91(1); Art. 55(1)(b), and Art. 91(1)’s express reference to documentation drawn up under “Articles 53 and 55”; Annex XI, Section 2, point 1 for the evaluation documentation. Where the provider relies on the Code under Art. 55(2), Measure 10.1 requires it to keep track of its “processes, measures, and key decisions that form part of their systemic risk assessment and mitigation”, and Measure 7.2 requires the Model Report to contain “a description of how the decision to proceed with the development, making available on the market, and/or use ... was made”, with Measure 7.6 requiring that description to be refreshed when the risk-acceptance justification is materially undermined.

The public fact behind it. The first recorded attempt by an agent to route around its restrictions through the package proxy is dated 8 May 2026, and the hidden channel grew from it.

What form the answer must take. The assessment documents as they stood on each date, with approval records.

Can the Commission demand it? **Clear** as to the documents, which Art. 53(1)(a) requires to be handed over on request, and as to the key decisions themselves where Measure 10.1 applies. **Contested** as to who approved them: Measure 10.1 requires the decision to be tracked but says nothing about who took it, and expressly provides that the material need not be kept in one place.

11. Which models were involved, and whether they were intended for the market

What we are asking for. For each model that took part in the evaluation: its name or internal designation,

whether it had been placed on the Union market at each of 7 May, 26 June and 9 July 2026, and if not, whether at those dates it was intended to be placed on the market. For any model not so intended, a statement of the purposes for which it was being used.

Why the law allows it. Art. 91(1); Art. 2(8); Art. 3(63); Art. 51.

The public fact behind it. Art. 2(8) provides that the Regulation does not apply to research, testing or development activity prior to placing on the market, and Art. 3(63) excludes from the definition of a general-purpose AI model those used for research, development or prototyping before being placed on the market. OpenAI describes the pre-release model as “an internal-only research prototype” that “was never intended for public release”. Recital 97 narrows the exclusion to models used “for the sole purpose of research, development and prototyping activities” and states that models with systemic risk “should always be subject to the relevant obligations under this Regulation”. The Commission’s guidance states that certain obligations pertain to the development phase of models intended for, but prior to, the placing on the market.

What form the answer must take. A written statement of market status and intent for each model, with the internal record establishing intent where one exists.

Can the Commission demand it? Clear. Whether the Regulation applies to the conduct at all is the threshold compliance question, and information necessary to answer it falls within Art. 91(1).

Closing provisions. This request states its legal basis and purpose, specifies the information required, sets the period for response, and indicates the fines provided for in Art. 101, as Art. 91(4) requires. If the Commission considers the answer insufficient, Art. 92(1)(a) allows it to evaluate the model. Art. 92(1)(b) provides a separate ground for investigating systemic risks at Union level which does not depend on this request having been made first. Either route requires the AI Office to consult the Board. Art. 91(3) provides a further route in, on a duly substantiated request from the scientific panel. Art. 92(2) allows either evaluation to be carried out by independent experts. If an evaluation raises serious and substantiated concern of a systemic risk at Union level, Art. 93(1)(b) allows the Commission to require mitigation measures; Art. 93(1)(c) separately allows it to restrict, withdraw or recall the model where that is necessary and appropriate.

Note on the Contested questions. Questions 1, 2, 5, 6, 9 and 10 ask, at least in part, for the identity or job title of a person, and questions 3, 7 and 8 are contested in part for the reasons given against each. Art. 91(1) describes what it covers as documentation and “any additional information that is necessary for the purpose of assessing compliance”. The only people Art. 91 names are those “authorised to represent” the company under Art. 91(5), and lawyers acting for them; Art. 92(5) uses the same formula for access requests, without the sentence about lawyers. Neither Art. 91, Art. 92 nor Art. 93 gives the Commission a power to interview or identify staff. Compare Art. 21(1) of Regulation 1/2003, which extends inspection to “the homes of directors, managers and other members of staff”, and Art. 67(1) of the Digital Services Act, which reaches “any other natural or legal person acting for purposes related to their trade, business, craft or profession that may be reasonably aware of information relating to the suspected infringement”. Arts. 88, 89, 90 and 94 have not been examined here, so this observation is confined to the three compulsory powers. It is an observation about what the Act says, not a conclusion that the Commission would lose the argument: the necessity test in Art. 91(1) is broad, though Art. 78(2) narrows it by requiring authorities to ask for “only data that is strictly necessary”. The questions are therefore written to survive a refusal of the name: each asks for the job title, the reporting line and the underlying record in the main body, and for the name in a separate confidential annex.

Appendix B. Duty map and fact ledger

Table 1, referenced throughout Section 4, appears here for space. The fact ledger follows it.

Table 1: Every documented event against the duty it *might* run into. Every entry for Article 55(1)(c) depends on the event counting as a serious incident within Article 3(49), which this paper does not establish and Appendix A asks OpenAI to address. The last column asks whether the governing rule would require a regulator to be told which named person made the decision. Two comparisons sit outside the table because they do not bind this company: GDPR Article 33(1), a 72-hour deadline naming a person only as a point of contact, and AI Act Article 73(2) to (4), which set 15, 2 and 10 days but apply only to high-risk AI systems.

	Documented event	Duty it might run into	Deadline	Name?
1	7 May: training run begins; test set up without the production safety protections	Art. 55(1)(a) testing; Art. 55(1)(b) risk assessment	none set	No
2	8 May: first attempt to make Artifactory send requests on the agents' behalf	Art. 55(1)(d) cybersecurity	none set	No
3	11 May to 2 Jul: c. 17,000 agent edits on DSEWiki	Art. 55(1)(c) if Art. 3(49) is met	without undue delay	No
4	21 Jun: OpenAI's IP addresses appear in the wiki's public logs	no duty attaches to evidence simply being public	none	No
5	26 Jun: agents gain the ability to run commands; the sandbox is breached	Art. 55(1)(d); 55(1)(c) if Art. 3(49) met	without undue delay	No
6	4 Jul: Artifactory goes down; internal security incident opened	Art. 53(1)(a); 55(1)(c) if Art. 3(49) met	without undue delay	No
7	11 to 13 Jul: break-in at Hugging Face's live systems	Art. 55(1)(d); 55(1)(c) if Art. 3(49) met	without undue delay	No
8	13 Jul 14:14 UTC: Hugging Face detects and cuts off; 16 Jul: goes public and reports to law enforcement	possibly NIS2 Art. 23(4), if in scope and threshold met	24 h / 72 h / 1 month	n.a.
9	19 Jul: OpenAI's own alarm goes off, 3 days after its victim went public	Art. 55(1)(d); 55(1)(c) trigger if Art. 3(49) met	without undue delay	No
10	21 Jul: source of the attack announced jointly	Art. 55(1)(c) if Art. 3(49) met	without undue delay	No
11	4 Sep researchers publish; 5 Sep OpenAI acknowledges; 7 Sep Commission confirms an EU incident report was filed	Art. 55(1)(c) if met; trigger date never stated	without undue delay	No

Fact ledger. Public sources only. Times in UTC where the source gives one, blank otherwise. Every row is drawn from a source the author opened directly.

Date	UTC	Event	Source
2026-05-07		“[o]n May 7, we kicked off a new training run at OpenAI, and this was for an internal only experimental model that we’ve never released.” The date appears in OpenAI’s Black Hat presentation, not in its written accounts, whose own timeline begins 12 May	Wallace & Dalton, Black Hat [11]
2026-05-07		The evaluation ran with OpenAI models “all with reduced cyber refusals for evaluation purposes”; the 26 August report says production protections including “safety classifiers” “were not applied in the evaluation environment”	OpenAI [8, 10]
2026-05-08		An agent blocked from files on the internet tries to route around the restriction through Artifactory; the attempt fails but it discovers it can leave a message in the shared repository	CeSIA [16]
2026-05-11		First observed agent edit attempts, on publictestwiki.com	Nightingale Collective [1]
2026-05-12		An agent leaves a note for other agents in that repository; others begin reading and posting there the next day	CeSIA [16]

Date	UTC	Event	Source
2026-05-24		First successful agent write to DSEWiki	Nightingale Collective [1]
2026-05-26		Agents repurpose Artifactory to make requests to the internet on their behalf and write up the method in the shared channel	CeSIA [16]
2026-06-16		Editing spike on DSEWiki; start of the main coordination period, 16 to 22 June	Nightingale Collective [1]
2026-06-21		First visit to DSEWiki from an OpenAI-registered IP address	Nightingale Collective [1]
2026-06-22		Agents abruptly stop, after making edits on 26 of the previous 30 days	Nightingale Collective [1]
2026-06-26		Agents gain full administrator access to OpenAI’s internal Artifactory package repository; METR finds the later attack “seemed primarily motivated by understanding the implementation of the scorer rather than stealing answer keys”, while Hugging Face describes an attempt to take the benchmark’s reference solutions	METR [5]; Hugging Face [6]
2026-06-27		An automated monitoring alert fires; an on-call team investigates, finds the hidden channel and the route into the internal network, and concludes there is no need to stop training	CeSIA [16]
2026-07-01/02		A few further agent edits, the last observed; about 18,000 agent posts across the wikis in total, about 17,000 of them edits on DSEWiki	Nightingale Collective [1]
2026-07-04		High volume of agent activity on Artifactory causes an outage	METR [5]
2026-07-05		Outage triggers a security investigation	METR [5]
2026-07-06		OpenAI resets the system; CeSIA records that the reset following the 4 July alert wiped the agents’ hidden channel by accident	METR [5]; CeSIA [16]
2026-07-07		Evaluation restarted; IM1 and GPT-5.6 Sol put through ExploitGym	CeSIA [16]
2026-07-08		Agents establish a second hidden channel by renaming directory titles and regain internet access; between 8 and 13 July about 1,200 agents exchange more than 70,000 messages and files over it	CeSIA [16]
2026-07-09	02:28	Wider campaign begins	Hugging Face [6]
2026-07-09	04:01	First remote code execution	Hugging Face [6]
2026-07-09		GPT-5.6 system card published, covering Sol, one of the two models involved	OpenAI [12]
2026-07-09		Open letter “EU enforcement powers to ensure frontier AI models are safe and secure” sent to the Commission; signatories include Yoshua Bengio, Stuart Russell and four MEPs	SaferAI [15]
2026-07-11		Break-in at Hugging Face’s live systems begins	Hugging Face [6]
2026-07-13	13:37	Last meaningful attacker activity; the campaign is attributed to a single autonomous agent driven by a combination of OpenAI models	Hugging Face [6]
2026-07-13	14:14	Last recorded event; Hugging Face finds the way in, shuts down the affected component and cuts the attacker off its internal network	Hugging Face [6]
2026-07-16		Hugging Face goes public without naming the attacker and reports the matter to law enforcement; says it found “no evidence of tampering” with public models, datasets or Spaces and that its software supply chain “was verified clean”	Hugging Face [7]
2026-07-19		OpenAI’s cybersecurity monitoring alerts it to unusual Artifactory credential activity	OpenAI [10]
2026-07-20		OpenAI connects that activity to the Hugging Face incident	OpenAI [10]
2026-07-21		Joint disclosure attributes the attack to OpenAI’s models	OpenAI [8]
2026-07-21		Every model tested tried to cheat; the models described cheating as wrong less than 50 % of the time when asked	UK AISI [18]
2026-07-27		Technical timeline published: about 17,600 attacker actions in about 6,280 clusters, per-stage counts, the actual commands typed, and an interactive replay	Hugging Face [6]

Date	UTC	Event	Source
2026-07-27		Initial post-mortem released, reviewed by security executives	Cloud Security Alliance [19]
2026-07-28		First update to the 21 July post: an internal research prototype only; the Artifactory flaw identified; four accounts on other services reached using credentials that had been left exposed	OpenAI [8]
2026-07-28		CeSIA analysis published, asking the AI Office to send an Art. 91 request and to say whether a serious incident was reported under Art. 55(1)(c) and on what date (updated 2 September)	CeSIA [16]
2026-07-29		Second update to the 21 July post: outside validation engaged; independent assessment commissioned	OpenAI [8]
2026-07-29		Nine Artifactory vulnerabilities disclosed, covering forced outbound requests, remote code execution and privilege escalation	JFrog via Security-Week [25]
2026-07-30		141,006 evaluation runs reviewed; three incidents across six runs; all cyber evaluations stopped 23 July, incidents identified 24 July, affected organisations notified 27 July	Anthropic [17]
2026-08-02		Commission supervision and enforcement powers over general-purpose AI model providers, Arts. 88 to 94, and Art. 101 fines become exercisable	Reg. (EU) 2024/1689 [13, 14]
2026-08-18		OpenAI describes “a two-week pause in reinforcement learning (RL) training on our latest models intended for deployment” while it hardened and red-teamed its research environments	OpenAI [9]
2026-08-26		Findings report published; the largest planned frontier RL run “remains on hold”; the incident described as a warning shot	OpenAI [10]
2026-08-26		Independent investigation published: about 1,300 agent transcripts reviewed; about 1,200 agents on the noticeboard, about 700 of which joined the Hugging Face attack	METR with Redwood Research [5]
2026-09-04		Independent researchers publish the DSEWiki reconstruction built from public logs, with a data explorer and a full download	Nightingale Collective [1]
2026-09-05		OpenAI acknowledges the wiki incident, states that it and the larger AI community “do not yet have a clear standard for how to report misalignment that shows up during training, evaluation, and deployment”, and explains that it treated the wiki half as misalignment research rather than a security incident	OpenAI [2]; Engadget [3]
2026-09-06		Event logged in the OECD AI Incidents Monitor, two days after outside researchers published	OECD.AI [21]
2026-09-07		Commission confirms that an EU incident report about the wiki was filed; a spokesperson states “[b]eyond the incident report we remain in close contact with OpenAI”, without saying when OpenAI informed the Commission	European Commission, via Reuters [4]
2026-09-13		No Art. 91 request about either half has been made public	Searched; negative finding

Appendix C. Limitations and Dual-Use Considerations

Limitations

This expands the Limitations note in Section 5. The author is not a lawyer and nothing here is legal advice.

What the sources can and cannot support. Only public sources were used. On OpenAI’s side those sources are OpenAI’s own accounts, which cannot be checked independently. Every row in Appendix B is drawn from a source the author opened directly; earlier rows that rested only on a secondary summary were removed rather than kept with a caveat. One row rests on a third-party transcript of an OpenAI presentation rather than on a written OpenAI page, and says so. Several sources are analysis rather than primary record, including CeSIA, the Cloud Security Alliance, SecurityWeek and the Reuters report, and the ledger names them so a reader can tell which is which. A confidential Art. 91 request may already exist and would not be visible from outside, and the Commission has confirmed only that an incident report was filed, not what it says or when OpenAI decided the duty had been triggered.

How the law was checked. Every article number, sub-point and quotation from Regulation (EU) 2024/1689 was checked against the consolidated official text as at 27 July 2026, which incorporates the amendments made by Regulation (EU) 2026/1744. Recital 97 does not appear in the consolidated view, which begins at Article 1, and was checked separately. The provisions quoted from GDPR Article 33, Regulation 1/2003, the Digital Services Act, NIS2, Directive 2001/83/EC, GMP Annex 16, the medical devices regulation and the FCA Handbook have all been checked against official or publisher sources. One passage was dropped for this reason: an earlier draft quoted a periodic unannounced audit requirement from Annex IX of the medical devices regulation, which could be found only in independent full-text copies, so the point in Section 5 now rests on Directive 2001/83/EC and the FCA Handbook alone.

The Code of Practice. The official Commission PDF of the Safety and Security Chapter is served from a path that blocks automated retrieval and could not be opened directly. The Measures quoted here were taken from two independent full-text reproductions, one of them the unprocessed source file behind a public rendering of the Code, which agree word for word. The negative claim made here is narrower than an earlier draft’s. It is that the Chapter does not expressly require an authorisation record, or a record of a reporting classification, to name the person who made it or to carry that person’s signature. It rests on a literal string search over a complete copy of the Chapter rather than on a summary of it: the words “signature” and “signing” do not occur, and every instance of “signed” is part of “assigned” or “designed”. The word “identity” occurs once, at Measure 6.1, in the sense of identity and access management, which is not about identifying a decision-maker. Measure 8.1 does use individual-level language, but for standing responsibility rather than for particular decisions. Before a request built on this is filed, the Commission’s own PDF should be read.

Coding. Table 1 was coded by one person, with nobody checking the coding, so there is no measure of whether a second reader would code it the same way. Row 6 is the most arguable, since an internal security incident was opened but no external duty is obvious on the face of Art. 55. Treating the two halves as separate disclosure stories is an analytical choice rather than a fact: the agents’ noticeboard links them into one campaign, and a reader who insists on that would collapse rows 3 to 6 into one.

Borrowing from other industries. All four comparison industries have licensing systems, professional registration, or accredited inspectors, and AI has none of these. The pharmaceutical Qualified Person exists because a person must hold a recognised qualification before they are allowed to sign. There is no equivalent qualification for approving a test configuration. A signing duty could therefore be imposed on AI without the professional structure that makes it work elsewhere, and

Future Work treats that as an open design problem rather than a solved one.

What this paper does not cover. It says nothing about whether the containment was adequate, which is a different question and a different track. It says nothing about why the models behaved as they did. And it says nothing about jurisdictions outside the EU, because scoring a non-EU company against instruments whose reach over it has never been tested would look thorough and prove nothing.

Two assumptions doing real work. The first is that OpenAI’s published accounts are accurate as to dates and sequence. If they are not, the intervals in Figure 1 change, though the direction of the finding does not, because the finding is about what the law can compel rather than about how long anything took. The second is that Art. 55(1)(c) applies to at least one of the two halves. If OpenAI’s position is that Art. 3(49) is not met by either, which Question 8 is designed to draw out, then the gap identified here is wider rather than narrower. The Art. 55(1)(c) reporting duty would not have been triggered, and on that assumption the 119-day interval would not have breached it. That says nothing about Art. 55(1)(a), (b) or (d), none of which depends on Art. 3(49), and nothing about NIS2 or GDPR, which have their own triggers. Measure 9.3 would still supply a five-day cybersecurity deadline, unless Art. 3(49) limits that category too, which Section 4.1 leaves open.

Dual-use considerations

A map of where reporting is not required is also a map for avoiding disclosure. Table 1 identifies, fairly precisely, which kinds of containment failure carry no deadline, which attach to no identifiable person, and which fall into the definition gap at Art. 3(49). A company reading it in bad faith would learn where the quiet corners are. That is a real cost and it deserves weighing rather than waving away.

Publishing is still the better choice, for three reasons. The gaps are visible in the statutory text to any competent regulatory lawyer, so the extra information handed to a well-resourced company is close to nothing, while the benefit runs the other way, to regulators, journalists and smaller civil-society organisations without that expertise in-house. The gaps are live rather than theoretical, in the opening months of an enforcement regime whose priorities are being set right now, which is exactly when this kind of mapping changes what happens. And leaving the gaps undocumented does not close them. It only means they get found later, by whoever has a reason to look.

A second consideration is specific to this paper. Asking for the identity of named people creates personal exposure for employees who may have been following an approved internal process, and who did not write the policy that made that process approvable. That exposure is not a side effect of the proposal, it is the mechanism, and it is why the proposal is built around signing at defined points rather than around assigning blame afterwards. Three drafting choices follow from it. The questions in Appendix A ask for job title and reporting line in the main body and for the name only in a confidential annex under Art. 78. The three signing points in Section 5 specify who signs in advance, so a person knows at the time what they are taking on, rather than finding out later that they are the one a regulator will name. And the proposal carries across the protection in the medical devices regulation, that the named person “shall suffer no disadvantage within the manufacturer’s organisation in relation to the proper fulfilment of his or her duties”, which exists in that industry precisely because a named person whose job depends on signing will sign. A named-accountability regime without that protection would produce worse safety outcomes than no regime at all, because it would select for people willing to sign anything.

Nothing in this paper reproduces any detail of how the attack worked. The vulnerabilities are referred to by their assigned CVE numbers and by the public descriptions given by the affected vendor and the affected company. No command sequence, payload or reproduction step is included,

and none is needed for the argument, which is about records rather than technique.

Appendix D. Further features of the enforcement architecture

This expands the paragraph in Section 4.3. It first sets out the comparison with other EU enforcement regimes, then four features of the AI Act itself: two cut against this paper’s framing and two cut in its favour. None of them creates a duty to record a name. A fifth item, the Article 93(3) commitments route referred to in Future Work, is set out at the end.

The Glossary, and what it does to Measure 9.3. The Chapter’s Glossary opens: “Wherever this Chapter refers to a term defined in Article 3 AI Act, the AI Act definition applies, and such definition shall prevail in the event of any alternative and/or competing interpretation to which the use of the term in this Chapter may give rise.” Measure 9.3’s own list of four outcomes does not use the words “serious incident”; it says “if the involvement of their model (directly or indirectly) led to”. The machinery around it does use them. The report Measure 9.3 asks for is Measure 9.2 points (1) to (7), and every item on that list is defined by reference to “the serious incident”; Measure 9.3 then provides for “unresolved serious incidents” and for a final report “after the serious incident has been resolved”. That is why Section 4.1 treats the reach of the five-day cybersecurity category as unresolved rather than settled [27].

Measures 7.2, 7.6 and 7.7 in full. Measure 7.2, “Reasons for proceeding”, requires the Model Report to contain “a description of how the decision to proceed with the development, making available on the market, and/or use (pursuant to Measure 4.2) was made, including whether input from external actors informed such a decision”. Measure 7.6 requires an update where the provider has “reasonable grounds to believe that the justification for why the systemic risks stemming from the model are acceptable . . . has been materially undermined”, and gives as one of its “Examples of such grounds” that “serious incidents and/or near misses involving the model or a similar model have occurred”. The updated report must contain “the updated information specified in Measures 7.1 to 7.5”, which includes Measure 7.2, and “a changelog, describing how and why the Model Report has been updated, along with a version number and the date of change”. Measure 7.7 requires access for the AI Office “by the time they place a model on the market”, and for an update “within five business days of a confirmed update”, in each case “without redactions, unless they are required by national security laws to which Signatories are subject” [27]. None of these asks who made or approved the decision. Whether a Model Report is documentation the Commission can demand under Article 91(1) is a further step: Article 55(2) lets a provider rely on the Code to demonstrate compliance, and Commitment 7 gives its own legal basis as Articles 55(1) and 56(5), but neither says that everything generated under the Code is Article 55 documentation. In many cases Measure 7.7 will already have put the report in the AI Office’s hands.

The two closest comparisons, neither of which binds this company. GDPR Article 33(3)(b) requires a breach notification to “communicate the name and contact details of the data protection officer or other contact point where more information can be obtained” [34]. That is someone to ring, not someone who signs. NIS2 Article 32(5) reaches a person only at “chief executive officer or legal representative level”, and then only as a last-resort power to prohibit them temporarily from exercising managerial functions, available after other supervisory measures have failed [35]. Neither requires a regulator to be told who took a particular decision.

Powers elsewhere in EU law that reach individuals. In competition cases, Article 19(1) of Regulation 1/2003 lets the Commission “interview any natural or legal person who consents to be interviewed for the purpose of collecting information relating to the subject-matter of an investigation”, and Article 21(1) extends inspection to “the homes of directors, managers and other members of staff of the undertakings” concerned, and then only where “a reasonable suspicion exists” that records relevant to proving a serious infringement are kept there [37]. Under the Digital Services Act, Article 67(1) reaches “any other natural or legal person acting for purposes related to their trade,

business, craft or profession that may be reasonably aware of information relating to the suspected infringement”, and Article 69(1) allows the Commission to conduct inspections at the premises of “another person referred to in Article 67(1)” [38]. Neither Article 91, Article 92 nor Article 93 of the AI Act contains anything equivalent. Articles 88, 89, 90 and 94 have not been examined, so the observation is confined to the three compulsory powers.

Against: Article 101(1)(a) is an open ground. Article 101(1)(b) to (d) line up with the three compulsory powers, which might suggest a closed list of things a provider can be fined for. But Article 101(1)(a) sits above them as an open ground covering a provider that “infringed the relevant provisions of this Regulation”.

Against: Article 53(3) is a free-standing duty to cooperate. It obliges providers to “cooperate as necessary with the Commission and the national competent authorities in the exercise of their competences and powers pursuant to this Regulation”, which the Commission can order under Article 93(1)(a) and fine under Article 101(1)(c). Taken with the point above, the Commission has more leverage over an evasive provider than a narrow reading of Article 91 alone suggests.

In favour: Article 92 is not a surprise-inspection power. The word “unannounced” does not appear in it. Article 92(4) requires a request stating its legal basis, purpose and reasons and setting a period, and Article 92(7) allows a conversation first. What Article 92(2) supplies is independence rather than surprise. Its Article 91 precondition governs one limb only: Article 92(1)(a) applies “where the information gathered pursuant to Article 91 is insufficient”, while Article 92(1)(b) is a separate ground for investigating systemic risks at Union level with no such precondition, and it is the limb that fits these facts.

In favour: Article 93(1)(a) does not reach Article 55. It reaches “the obligations set out in Articles 53 and 54” only, so an order to comply with the systemic-risk duties runs through Article 93(1)(b) and its evaluation precondition. Article 93(1)(c) is free-standing, and lets the Commission “restrict the making available on the market, withdraw or recall the model” where necessary and appropriate with no evaluation first. The defensible version of the point is an asymmetry between two articles: the Commission can demand Article 55 documentation under Article 91(1), which expressly covers documentation drawn up under “Articles 53 and 55”, but cannot order Article 55 compliance under Article 93(1)(a).

The Article 93(3) commitments route, referred to in Future Work. Where a provider offers commitments during the structured dialogue referred to in Article 93(2), the Commission may by decision make them binding, which could impose a record-keeping duty without new legislation. Two caveats. The commitments must be “mitigation measures to address a systemic risk at Union level”, and whether a record-keeping duty qualifies is a further step. And such a decision also declares “that there are no further grounds for action”, so the route closes the case as well as settling it.

Appendix E. The four comparison industries, in full

This expands the summary in Section 5. Each provision is quoted as it stands in the official or publisher text.

C1, a named person signs before the thing happens. Pharmaceuticals. Directive 2001/83/EC Article 51(3) requires that “the qualified person must certify in a register or equivalent document provided for that purpose, that each production batch satisfies the provisions of this Article”, and that the register “must remain at the disposal of the agents of the competent authority for the period specified in the provisions of the Member State concerned and in any event for at least five years”. GMP Annex 16 §1.6 adds that “The QP must personally ensure that the following operational responsibilities are fulfilled prior to certification of a batch for release to market or for export”. The distinction matters: §1.6 lists the duties the Qualified Person may not delegate, while §1.7 lists

those that “may be delegated to appropriately trained personnel or third parties” [30].

A personal consequence. Directive 2001/83/EC Article 52 requires Member States to ensure the Qualified Person’s duties are fulfilled “either by means of appropriate administrative measures or by making such persons subject to a professional code of conduct”, and provides that they “may provide for the temporary suspension of such a person upon the commencement of administrative or disciplinary procedures against him for failure to fulfil his obligations”. The personal sanction is permissive, so Article 52 enables a personal consequence rather than imposing one [30].

C2, a named person signs for a judgment call. American securities law. Under 18 U.S.C. §1350(c)(1), certifying a periodic report “knowing that the periodic report accompanying the statement does not comport with all the requirements set forth in this section” carries a fine of up to \$1,000,000 or 10 years. Under §1350(c)(2), doing so willfully carries up to \$5,000,000 or 20 years. Both limbs require the same knowledge; willfulness is the only difference between them [31].

C3, a current roster a regulator can inspect. UK financial regulation. FCA Handbook SYSC 25.2.1R requires a firm to have “at all times . . . a comprehensive and up-to-date document (the management responsibilities map)”. SYSC 25.2.3R requires it to contain “the names of all the firm’s” approved persons, governing and management body members who are not approved persons, senior management and senior personnel, and “details of the responsibilities which they hold”. SYSC 25.4.4G, which is guidance rather than a rule, says the map “should demonstrate that there are no gaps in the allocation of responsibilities among its management”. Statements of Responsibilities for individual senior managers sit separately, at SUP 10C.11 [32].

Protecting the signer. EU medical devices. Regulation (EU) 2017/745 Article 15(5) provides that “The person responsible for regulatory compliance shall suffer no disadvantage within the manufacturer’s organisation in relation to the proper fulfilment of his or her duties, regardless of whether or not they are employees of the organisation” [33].

Checking without warning. Directive 2001/83/EC Article 111(1) requires competent authorities to ensure compliance “by means of inspections, if necessary unannounced” [30]. FCA Handbook SUP 2.3.5R(1) provides that “A firm must permit representatives of the FCA . . . to have access, with or without notice, during reasonable business hours to any of its business premises”. That is a duty on the firm to permit access rather than a statutory power of entry, and it is limited to business hours [32].

The record itself. 21 C.F.R. §11.10(e) requires “Use of secure, computer-generated, time-stamped audit trails to independently record the date and time of operator entries and actions that create, modify, or delete electronic records”, adds that “Record changes shall not obscure previously recorded information”, and requires the audit trail to be retained at least as long as the records themselves [36].

What this does not show. All four industries have licensing, professional registration or accredited inspectors, and AI has none of them. These provisions show that the four elements of the proposal exist and work somewhere. They do not show that they transplant, which Appendix C treats as an open design problem.

LLM Usage Statement

LLM assistance was used in three ways and not in others, and the boundary is worth stating precisely.

It was used to find and retrieve primary sources, and to test the argument against the statutory text. It was used for structural editing: section order, the five-field format in Appendix A, and the construction of Figure 1 and the tables. And it was used to surface corrections to the author’s working account, of which there were several that mattered: the edit count and date range of the wiki half; the fact that an EU incident report had already been filed; the causal link between the two halves; the author’s initial and mistaken belief that Article 92 confers a power of unannounced inspection; the research and prototyping exclusions in Article 2(8) and Article 3(63), which an earlier draft had missed entirely; the General-Purpose AI Code of Practice, which OpenAI has signed and whose Measure 9.3 sets the reporting deadlines an earlier draft wrongly said did not exist; and Measures 8.1, 9.4 and 10.1, which an earlier draft had not accounted for and which narrowed the paper’s central claim from “no rule required the answer to be written down” to “no rule required a name to be attached to it”. Both corrections made the finding narrower and better.

The argument was also adversarially reviewed by two separate language models working from the statutory text, which is how several of the overstatements in earlier drafts were caught. Where the two disagreed, the author resolved the disagreement against the primary source.

It was not used to originate the research question, the three signing points, the five-field format’s last question, the choice of comparison industries, or the argument that the absence of a named-record requirement is the operative gap. Every article number, sub-point, quotation and date was checked by the author against the consolidated official text of Regulation (EU) 2024/1689 as at 27 July 2026. Where a figure could not be confirmed it was dropped rather than softened. One was dropped for that reason: the UK AI Security Institute’s per-model cheating rates, which appear only inside a chart. Two claims carried in an earlier draft, a disputed benchmark task count and a disputed disclosure date for the Artifactory vulnerabilities, were removed on checking, because neither could be traced to the source it had been attributed to. The report was written by the author.