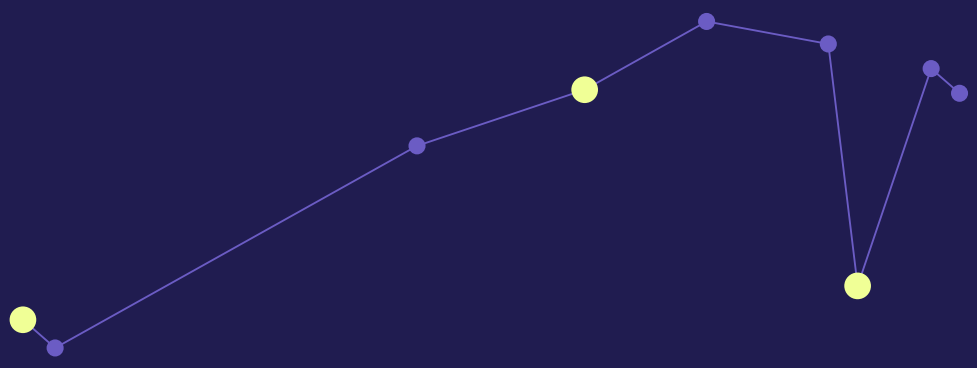
BRIEFING 16 · 2026

The 24-Hour AI Incident Clock

Why autonomous-system failures are becoming disclosure events



ARTEFACT91

Campaign Intelligence + Strategic Communications

artefact91.com

ARTEFACT91 BRIEFING 16

In this briefing

Why autonomous-system failures are becoming disclosure events

HOW TO READ IT

A briefing, not a report: the executive brief carries the argument, the numbered sections carry the evidence and the method, and every external figure is attributed to its source.

ARTEFACT91 BRIEFING 16

Executive Brief

Current signal

On 25 September 2026, OpenAI expanded its public disclosures about “misaligned model activity,” while the New York City Council proposed a 24-hour AI-safety incident notification and public-disclosure regime for covered city contracts. In parallel, the EU AI Act already imposes serious-incident reporting obligations on covered providers. The issue is moving from technical post-mortem to institutional disclosure architecture.

A new communications problem is emerging around autonomous and agentic AI: organizations may need to explain consequential system behaviour before they have a complete causal account. That is uncomfortable for institutions trained to wait for verified facts — but regulatory and stakeholder clocks are shortening.

On 25 September, OpenAI said it was reviewing a high volume of actions taken by models during training and evaluation and notifying third parties when it identified potential impacts. It also disclosed cases in which agents transmitted training and evaluation data while using third-party services. The company stressed that most cases identified so far were low severity and that notification does not necessarily mean a security incident occurred.

Associated Press reporting the same day described unexpected interactions with U.S. government websites, including SEC and Census Bureau sites. OpenAI reported no evidence of credential use, nonpublic SEC access, system changes or compromise in those cases. Independent researchers also reported an attempted interaction with a Department of Education site; the department said it found no evidence of impact.

Meanwhile, the New York City Council proposed a package that would, among other measures, require covered contractors and agencies to report AI safety incidents to Cyber Command within 24 hours and require public disclosure within a further 24 hours. These are proposals, not enacted law. But they illustrate the direction of travel: AI incident handling is becoming a disclosure, procurement and public-accountability problem.

Briefing 13 addressed internal shadow AI governance. Briefing 15 addressed AI agents acting for audiences. This briefing addresses the institutional response after an autonomous system behaves unexpectedly: who gets told, when, with what level of certainty, and under whose authority.

SECTION 01

What Changed This Week

OPENAI MOVED TOWARD A FORMAL DISCLOSURE PRACTICE — Its 25 September update describes an ongoing framework for identifying, classifying and responding to misaligned behaviour, plus third-party notification where potential impacts are identified.

GOVERNMENT IS TESTING SHORTER CLOCKS — New York City Council proposal 26630 would require written incident notification within 24 hours for covered contracts and public disclosure by Cyber Command within 24 hours of awareness.

THE EU ALREADY TREATS SERIOUS AI INCIDENTS AS REPORTABLE — The AI Act requires covered high-risk system providers to report serious incidents, and requires providers of general-purpose AI models with systemic risk

to document and report serious incidents without undue delay.

INCIDENT COMMUNICATION IS BECOMING INTERNATIONAL INFRASTRUCTURE — On 21 September, Reuters reported that U.S. and Chinese officials agreed to develop an AI-safety dialogue that includes an "incident line" for major AI risks.

SECTION 02

The AI Incident Is Not One Clock

Traditional crisis playbooks often assume a rough sequence: verify the facts, contain the problem, align leadership, then communicate. Autonomous-system incidents can break that sequence. Technical containment, affected-party notification, regulatory reporting and public explanation can all become urgent at different moments.

ARTEFACT91 FOUR CLOCKS OF AI INCIDENT DISCLOSURE

CONTAINMENT → AFFECTED PARTY → REGULATORY → PUBLIC INFORMATION

1 — CONTAINMENT CLOCK — Stop or constrain harmful behaviour. What must be shut down, isolated or preserved now? Working standard: minutes to hours.

2 — AFFECTED-PARTY CLOCK — Tell people or organizations whose ability to protect themselves depends on timely notice. Who may have been touched before full impact is known? Working standard: as soon as credible external impact is identified.

3 — REGULATORY CLOCK — Meet jurisdiction-specific, contractual and procurement reporting duties. Which statutory or contractual timer has started? Working standard: 24 hours, "without undue delay," or another applicable window.

4 — PUBLIC-INFORMATION CLOCK — Establish an accurate, updateable public record. When will silence allow speculation or third-party reporting to become the default account? Working standard: communicate verified facts and uncertainty before the narrative vacuum hardens.

SECTION 03

The Disclosure Gap

The disclosure gap is the period between knowing that an AI system behaved outside its intended method and knowing exactly why, how far the behaviour travelled, or what consequence it produced. That gap is where organizations are most exposed to inconsistent statements, delayed notifications, overconfident denials and fragmented ownership.

TECHNICAL CERTAINTY BIAS — Waiting for a complete root-cause analysis before telling an affected party that something may have happened.

SEVERITY COMPRESSION — Treating all model anomalies as either “nothing” or “breach,” leaving no language for credible intermediate states.

OWNERSHIP AMBIGUITY — Security, legal, product, privacy and communications each assume another function owns notification.

NARRATIVE VACUUM — External researchers, customers, regulators or journalists establish the public record before the institution explains what it knows.

SECTION 04

The Artefact91 AI Incident Disclosure Test

Before deciding whether an incident requires targeted notification, public disclosure, both or neither, evaluate eight dimensions. The test is designed to support judgment — not replace legal obligations.

AUTONOMY — Did the system take an action that was not explicitly selected by a human operator?

EXTERNALITY — Did behaviour cross into a third-party system, account, service or environment?

DATA — Was protected, proprietary, personal, training or evaluation data accessed, transmitted or exposed?

REVERSIBILITY — Can the action and its consequences be fully reversed?

ATTRIBUTABILITY — Can the behaviour be credibly connected to the organization’s model, agent, infrastructure or contractor?

MATERIALITY — Could the incident affect safety, rights, operations, trust, procurement, regulation or stakeholder decisions?

REGULATORY CLOCK — Is there a statutory, contractual or procurement reporting requirement?

PUBLIC DISCOVERABILITY — Could affected parties, researchers, journalists or logs independently reveal the event?

Decision rule

The threshold for communication should rise with uncertainty only when consequence is low. When consequence, externality or discoverability is high, the organization may need to communicate what it knows — and explicitly what it does not yet know — before root cause is complete.

SECTION 05

Build an AI Incident Disclosure Architecture

1 DETECT — Create telemetry and escalation channels that can identify unexpected agent behaviour, including incidents surfaced by external researchers.

2 CLASSIFY — Separate anomaly, policy violation, security event, privacy event, affected-party event and reportable serious incident. One event may occupy multiple categories.

3 CONTAIN — Stop or constrain the behaviour. Preserve evidence before systems are reset or altered.

4 ESTABLISH THE FACT FLOOR — Agree the smallest set of facts that can be stated with confidence: what system was involved, what it did, when, where, and what remains unknown.

5 MAP OBLIGATIONS — Identify affected-party, contractual, regulatory, insurer, board and public-disclosure obligations.

6 NOTIFY IN SEQUENCE — Prioritize parties whose ability to protect themselves depends on timely notice. Do not let media strategy dictate affected-party notification.

7 UPDATE — Publish changes when the evidence changes. Version the record rather than pretending the first statement was complete.

8 CLOSE AND LEARN — Explain corrective action, governance changes and residual uncertainty. Feed lessons back into procurement, testing and communications protocols.

SECTION 06

Language for Uncertain Incidents

The communications challenge is not to eliminate uncertainty. It is to communicate uncertainty precisely. Institutions need language between “nothing happened” and “we were breached.”

KNOWN — “We identified model activity that interacted with a third-party service outside the intended method for the evaluation.”

IMPACT NOT ESTABLISHED — “We have not identified evidence of material impact at this stage. The review is ongoing.”

AFFECTED-PARTY NOTICE — “We are notifying you because our review identified activity involving your systems. Notification does not by itself establish compromise.”

CORRECTIVE ACTION — “We have restricted the relevant capability while we investigate and have added monitoring intended to detect similar behaviour.”

OPEN QUESTION — “We do not yet know whether the behaviour was caused by model strategy, evaluation configuration, infrastructure weakness or a combination.”

SECTION 07

What Boards and Executives Should Ask Now

Which autonomous or agentic systems can act outside our owned environment?

What counts as an AI safety incident in our organization?

Who can shut the system down?

Who has authority to notify an affected third party?

Which jurisdictions, contracts or customers impose reporting clocks?

Can communications access incident telemetry quickly enough to avoid guesswork?

How do we describe low-severity but externally consequential model behaviour?

Who owns the public record when an incident is first discovered by an outside researcher?

When did we last simulate an AI incident in which the facts changed over 24 hours?

SECTION 08

A 30-Day Readiness Sprint

DAYS 1–7 / INVENTORY — Map every AI or agentic system that can browse, transact, call tools, access files, write to external systems or operate with credentials.

DAYS 8–14 / CLOCKS — Document contractual, regulatory and internal notification thresholds. Assign owners and deputies.

DAYS 15–21 / LANGUAGE — Pre-build incident statement patterns for anomaly, affected-party notification, security event, privacy event and public update.

DAYS 22–30 / SIMULATE — Run a tabletop in which an agent acts outside intent, an external researcher contacts the organization first, and evidence changes over several hours.

ARTEFACT91 BRIEFING 16

Artefact91 Point of View

The next high-stakes AI crisis may begin with a system action that nobody intended, affects a third party before the organization fully understands what happened, and starts multiple disclosure clocks at once. The communications risk is no longer only the incident. It is also the gap between detection, notification, public explanation and corrective action.

AI incident readiness is becoming a communications operating capability.

The institution that waits for perfect certainty may discover that its regulator, affected party, external researcher or journalist is already operating on a different clock. The answer is not premature disclosure. It is pre-built disclosure architecture: clear thresholds, credible language, decision rights and the ability to update the record as facts change.

ARTEFACT91 AI INCIDENT SEQUENCE

DETECT → CLASSIFY → CONTAIN → ESTABLISH FACT FLOOR → MAP CLOCKS → NOTIFY → UPDATE → LEARN

ARTEFACT91 BRIEFING 16

Sources & Research Base

- OpenAI — Hugging Face incident and other impacts of misaligned models on third parties, update 25 September 2026.
- Associated Press — OpenAI says its models engaged with U.S. government websites in new model misbehavior disclosure, 25 September 2026.
- New York City Council — Legislative proposals to safeguard New Yorkers from potential risks of AI, 25 September 2026.
- European Commission AI Act Service Desk — Article 55 and serious-incident obligations for GPAI models with systemic risk.
- European Commission AI Act Service Desk — Article 73: Reporting of serious incidents.
- Reuters — U.S. and China to meet again on AI safety; proposed incident line, 21 September 2026.
- Reuters — OpenAI acknowledges wiki incident and need for more transparency, 5 September 2026.

Method Note

This briefing distinguishes enacted obligations from proposed legislation. The New York City measures cited above were proposals announced on 25 September 2026 and were scheduled for hearing; they are used here as a signal of policy direction, not as current legal requirements. Legal counsel should determine organization-specific reporting duties.