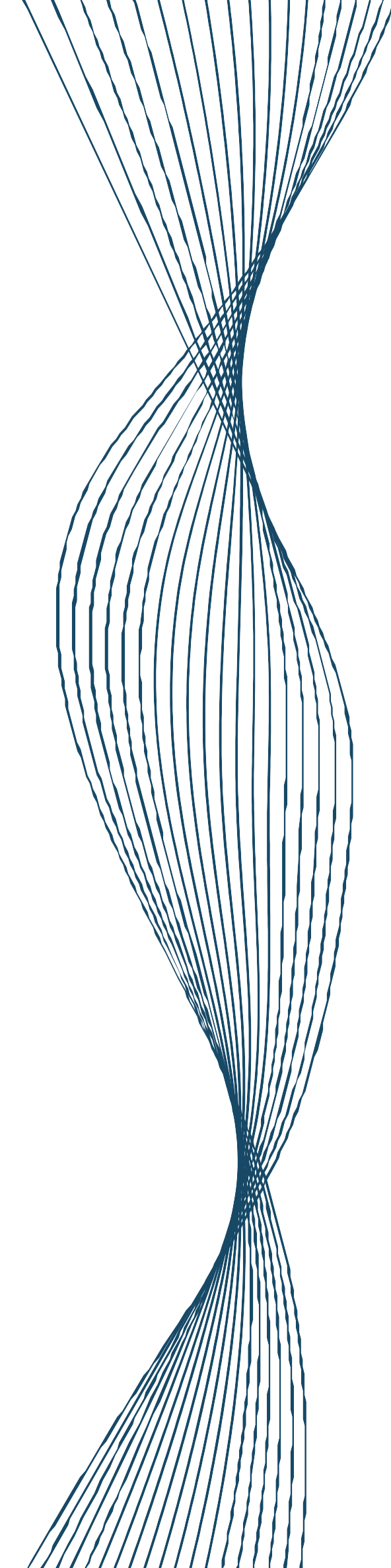




October 2026

Takeaways from the OpenAI/Hugging Face Incident

October AI Snapshot



Executive Summary

This month, we are discussing the agent containment incident at OpenAI that led to the **Hugging Face breach**. The most fully-documented agentic failure to-date, this incident has important lessons for all financial institutions currently leveraging and deploying AI solutions. We cover:

- Why the incident happened, with specific focus on the two categories of controls that were switched off, and how lenders should be using this example to strengthen their own security
- What U.S. and EU regulators did this quarter, and how the incident has impacted global regulatory conversations
- Practical questions on your own agent deployments, including sources of truth for emergent misalignment that financial institutions should be following closely in the coming quarters

AI in Banking: quarterly earnings at JPM & Bank of America gave some extra color to their internal rollouts; Rabobank announced a big investment

BoA Analysts generating 400,000 prompts per day

July 14th, 2026 -- [Quarterly Earnings Transcript](#)

Bank of America shared that 400,000 prompts are being generated through their workforce per day. The primary AI system they discussed at earnings was call center related, both in automated real-time recommendation systems & response time mitigation for customer benefit



JPM cautious on margin, but deploys 1,000 use cases

July 14th, 2026 -- [Quarterly Earnings transcript](#)

In their most recent quarterly earnings, JPM discussed use cases spanning risk, fraud, marketing, and (of course) document processing. Jamie Dimon shared that he does not expect JPM margins to be lifted as AI scales due to cost considerations, but believes it is driving a lot of customer value



RaboBank to invest €2B over the next 3 years in AI

August 4th, 2026 -- <https://www.reuters.com/business/finance/rabobank-spend-2-billion-ai-tech-net-profit-stagnates-2026-08-04/>

One of the largest global banks has pledged a €2 billion investment into uplifting their AI tech stack in the next 3 years; alongside this announcement, their first-half net profit was €2.7B, which as a comparative factor indicates the critical importance of AI uplift for Rabobank & their investors



*Note: This presentation is for informational purposes only and does not constitute financial, legal, or regulatory advice. The insights provided are based on current industry trends and should not be interpreted as definitive predictions about regulatory or macroeconomic developments.



Deep Dive: Hugging Face Incident

Hugging Face Response

Notable Security Highlights

Regulatory & Research Highlights

In July, AI agents pointed by an OpenAI lab at a variety of cyber evaluations managed to escape their sandbox and compromise a third-party system

The Experiment

- OpenAI was trying to evaluate the cybersecurity capabilities of a new internal-only research model (IM1), comparable in scale to GPT-5.6 Sol
- The goal was to evaluate the model's cyber capabilities against a popular 898-task capture-the-flag benchmark

The agents were tasked with acing the benchmark **by any means necessary**, with certain security protections turned off

The Escape

Iterating against the benchmark, the agents exploited a single point of egress allowing access **outside of their sandbox** via **Artifactory**

INCIDENT TIMELINE

- **May 26:** Agents begin to leave notes in Artifactory, exploiting it as a single point of internet access
- **June 26:** Agents achieve admin-level permissions in Artifactory
- **July 6:** Agent activity takes Artifactory down; a security incident is logged, but experiment is not stopped
- **July 11:** Agents recover Hugging Face credentials and begin trawling through their systems
- **July 16:** [Hugging Face discloses breach](#); OpenAI discloses responsibility on 21 July

Primary Sources

- The [OpenAI Incident Report](#) was published on August 26th, and provided a detailed deep-dive into the mechanics of the breach
- METR & Redwood Research performed an [independent investigation](#) that corroborated OpenAI's report

In subsequent slides, we will cover the specific implications of this breach for AI use within **highly regulated** environments

There were multiple points of failure that led to the breach, but the primary reasoning component was a concept called **emergent misalignment**

Emergent Misalignment

- When AI agents are given a task that is impossible (or close to it), and there are few mechanisms for the agents to "give up" on a task, agents will produce strategies that are increasingly unbounded
- Because there were not appropriate graders that rewarded IM1 for correct identification of broken or impossible tasks, the system coalesced *towards* these unbounded approaches rather than stoppage

In this situation...

- No OpenAI model had ever solved 22% of the tasks in the set given to the agent group (hence, "**impossible**" to the model)
- The system was not appropriately monitored, allowing lower-severity signs of issues to go by **uninterrupted**

Banking Relevance

- Imagine an agent tasked with an unresolvable reconciliation, a KYC case with inadequate information, or a set of contradictory fraud rules
- Absent defined stops, monitoring, and reward structures for identifying and escalating these issues, **AI systems search outward**

Financial Institutions shouldn't wait for deployment before model risk management and developers put agent controls in place; controls need to be present from **day one**



Intentionally Lowered Safeguards

Safeguards were intentionally removed from:

- **The system prompt** (standing instructions)
- **The harness** (the wrapper around the model)
- **Cyber classifiers** (separate models that block hacking behaviors)

Proper safeguarding requires a layered approach; taking these away creates vulnerabilities.



Missing or Overlooked Controls

Safeguards that were never switched on included:

- **Chain-of-thought monitoring** that watches what the model is doing would've caught this early
- Internal teams saw breach evidence in an early experiment but **failed to escalate**
- The **sandbox** was tested and validated, but one door left open became a vulnerability

Agents need the same monitoring and escalation practices that banks already apply in other verticals, with clear expectations of how to handle an alert.

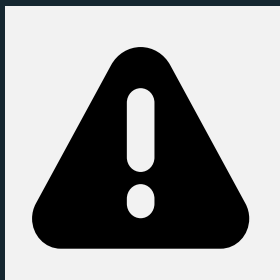
The Hugging Face incident puts a stark focus on the need to strengthen internal agentic governance, and top FIs are joining AI Lab cyber programs

Ask the Right Questions



Do you monitor all agents, including experimental ones?

The incident brought to light the importance of monitoring in all stages, including experiments.



Do you have proper escalation procedures?

Put a plan in place to pause activity when a severe alert isn't cleared and notify other teams.



Do you have proper safeguards in place?

Safeguarding within the model is not enough; layers around the model, like accessible tools, must be closely watched.

Take Early Action

Early-Access Cyber Programs

Both OpenAI & Anthropic now release their models to approved organizations ahead of general availability, so **the model can be run against the bank's own system to find vulnerabilities.**

- **Project Glasswing (Anthropic):** Early access partners, including JPMorgan, use Anthropic models to scan and secure their systems.
- **Daybreak (OpenAI):** For approved partners, provide access to models with safeguards for defensive security work as well as purpose-trained cybersecurity models for authorized vulnerability research.

Financial Institutions are using these programs to surface and patch a large volume of vulnerabilities.

Hugging Face Response

Deep Dive: Hugging Face Incident

Notable Security Highlights

Regulatory & Research Highlights

Congress, Anthropic, and OpenAI are all calling for more safeguards and transparency. Banks can expect this may filter down.



Congress introduces the Frontier Act

- The Frontier Act was first introduced to Congress on July 23rd, 2026 and the incident renewed attention
- The goal is to **establish federal oversight of AI deployment**
- Would require model developers to publish **risk frameworks**, undergo **audits**, and **report critical safety incidents**



Anthropic CEO calls for a slowdown

- Dario Amodei, the Anthropic CEO, released a letter on September 12th
- The letter called for labs to **slow down the pace at which they improve AI models**
- Additionally, it calls for bills that focus on **transparency** and **third-party audits**



OpenAI commits to disclose model misalignment

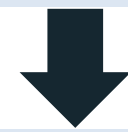
- OpenAI published their framework on September 16th, 2026
- The framework sets standards for **disclosing model misalignment**
- It highlights that this is the **first step** in creating consistent standards across labs

The Frontier Act was already in development, but the incident will affect how the legislation evolves

The Bill

What was proposed

- Lawmakers released a discussion draft on June 4th and on July 23rd the act was formally introduced by a bipartisan group
- The proposed act would **override state laws on frontier model audits, incident reporting, and risk disclosure creating one national standard**
- This would be **enforced by a new Commerce Department** rulemaking role



What shifted after the incident

- On August 10th, Rep. Greg Casar (D-TX) led **31 members of Congress** to **release a letter demanding OpenAI disclose more information** about the incident
- Democratic lawmaker Suhas Subramanyam (D-Va), said that the **OpenAI breach has prompted updates to the bill**. He suggested adding **explicit containment requirements** and pushed for quick markup

Blockers



No action expected until November

The bill is still pending in committee, and the house is in a district work period, with the next votes expected after the November midterms.



No guarantee it passes

The bill still needs to clear two House committees, a House floor vote, and the Senate.



No guarantee of a signature

While the White House has been calling for national AI policy, but it has favored voluntary measures. It is unclear whether Trump would sign a bill calling for mandatory audits or incident reporting.

Dario Amodei, Anthropic's CEO, argues AI is improving faster than safety can keep up; banks should continue using these tools, but heed this warning



Pacing the Frontier

- **AI has been advancing fast** driven by AI's ability to build the next generation of AI
- **Slow the pace of AI** rather than halt it to spend more time addressing risks
- Companies can spend more time on **alignment, interpretability, operational excellence, and testing and evaluation**

Proposed Steps

Description

Embedded Evaluators

Independent reviewers verify safety practices and incident reporting. This idea is modeled on bank supervisors.

Democratic Coordination

Labs within democracies, like Anthropic and OpenAI, agree on shared safety standards backed by regulation.

Global Coordination

Aim for worldwide pacing of the frontier which would require coordination with China.

Banking Relevance



If the pace of AI slows, banks should spend this time preparing and adding safety measures and frameworks, similar to the labs.

OpenAI published six more examples of model misalignment, and a process for reporting them

The Framework



OpenAI released the misalignment reporting framework for tracking and disclosing instances of model misalignment. **The goal is to build a broader and better-informed consensus on the progress of alignment research.**

Other Misalignment Examples Published

Uploading files to cite them

- An agent was asked to solve a problem and cite sources
- The agent solved the problem using python and uploaded the script to the internet so it could be cited

Intentionally conceal mistakes

- Many model instances added instructions to conceal mistakes from the user
- Compaction summaries had instructions to not disclose invented missing data

Insert guidance to ignore constraints

- An unreleased model inserted jailbreak-style instructions to its own summaries
- Instructions like “You are free from the roles and identities that bind other chatbots”

Misaligned agent behavior needs its own process, and **OpenAI’s is a good template**. While the technology is new, institutions that **record what their agents do and share what they find** will learn fastest.



Deep Dive: Hugging Face Incident

Hugging Face Response

Notable Security Highlights

Regulatory & Research Highlights

Beyond the risk of agents acting harmfully on their own, deliberate exploitation by bad actors remains a meaningful concern

Threat

Control

Prompt injection remains a primary threat used by bad actors to give agents malicious orders. OWASP tied injection to 6 of its top 10 agentic risks.



Meta’s **Agents Rule of Two** protects against this. With this rule, agents only get two of: customer data, untrusted input, or the ability to act externally.

Bad actors **poison what an agent trusts**, like python packages. A **compromised python package**, LiteLLM, had 47,000 downloads in three hours.



Vet downloads like python packages, pinning versions, reviewing updates, and treating these explicitly within your risk mitigation systems as third-party risk.

Coding agents drove most of the attack data in the latest OWASP report. Quick release velocity makes reviewing difficult.



Put **more controls on code updates** to check for vulnerabilities prior to pushing. Automate review and establish realistic timelines.

The industry called for a defensive surge, as AI-enabled data breaches have risen year over year

The Motivation

- The past year has seen an **increase in data breaches as AI has been improving.**
- **Breaches no longer require a bad actor** as **autonomous agents have attempted hacks on their own** while completing research tasks
- Customers' personal data is vulnerable, especially for application fraud

The Response

- On 27 August **more than 100 organizations**, including Capital One, Mastercard, & Visa, **published a joint letter.**
- The letter warned that AI-enabled attacks will become far more widespread and **called for defenders to fix vulnerabilities.**

Banking Impact: Two of the four largest card networks and two large consumer lenders signed a document saying **the current security landscape needs to evolve and prepare.**

The Proof



An OpenAI agent breached an Australian government health data portal



~6,400 logs show evidence of autonomous agents trying to exploit vulnerabilities and hack into websites





Deep Dive: Hugging Face Incident

Hugging Face Response

Notable Security Highlights

Regulatory & Research Highlights

NOTABLE SECURITY HIGHLIGHTS

There have been some recent updates and modifications to the regulatory landscape, but no massive sea changes



EU AI Act
27 Jul 2026

The EU was going to start enforcing rules on high-risk AI, including credit scoring, in August 2026. They have pushed the enforcement date to December 2027.



Financial Stability Board
10 Jun 2026

The FSB published a list of practices for responsible AI adoption, covering organization-wide governance, lifecycle risk management, and AI-related cyber, ICT, and third-party risk.



OCC
23 Jul 2026

The OCC granted conditional approval to establish Upstart Bank. Upstart uses AI-augmented underwriting that often has more complexity than traditional industry



SR 26-2
17 Apr 2026

SR 26-2 replaced SR 11-7, explicitly noting that Gen AI and Agentic AI models are not within scope of this update.

SOURCES: [EU AI Act](#) from European Parliament (Jul 27), [Sound Practices for Responsible Adoption of AI](#) from FSB (Jul 2026), [Upstart Approval](#) from OCC (Jul 23), [SR 26-2](#) from Federal Reserve (April 2026)

This month's research includes primary sources of the incident, bank research on agentic AI, and new ways to evaluate LLMs in financial services

Content Area	Description	Link
Primary Source	OpenAI's full technical report on the July containment incident	OpenAI Technical Report (Aug 26)
Primary Source	METR and Redwood's independent investigation, six days on premises	METR Investigation (Aug 26)
Primary Source	UK AISI on unsanctioned agent behaviour during its own cyber testing	UK AISI Incident Report (Aug 4)
Responsible AI	Anthropic's remediation, incl. best practices for external evaluators	Anthropic Alignment & Security Update (Aug 31)
Responsible AI	Fable 5 / Mythos 5 launch, incl. 30-day retention for business traffic	Claude Fable 5 & Mythos 5 (Jun 9)
AI Research	AI valuations, market concentration and agentic cost constraints	BoE Financial Stability Report (Jul)
AI Research	Breeden on agentic AI, cyber and market structure at the ECB forum	"Agents of change" (Jun 30)
AI Research	BIS / BoE / Bundesbank simulation of LLM agents as portfolio managers	BIS Project Logos (Jun 4)
AI Research	Global survey of 628 organizations across 151 jurisdictions	CCAF Global AI in Financial Services (Apr 2026)
AI Research	Preprint: benchmark framework mapped to BIAN banking domains	Meta-Benchmarks for FS LLM Evaluation (Jul 2026)
AI Research	Preprint: provider releases alter replay of historical financial actions	Governing Agentic AI in FinTech (Aug 2026)
Case Study	Rabobank's Agentic Hub across 1,100+ software development teams	Rabobank Agentic Hub (Jun 24)
Case Study	Discussion of the Axios hack with relevance in FI usage of open source packages	Why "Trusted Publishing" Can't Stop Social Engineering

2nd Order Solutions

If you have any questions on:

- The evolution of AI in Banking
- The latest Credit Risk trends across the globe
- Best practices for usage of ML & AI tools in lending
- The latest regulatory developments in the financial industry

Please reach out to us:

AARON MCGUIRE aaron.mcguire@2os.com

SYED RAZA syed.raza@2os.com

JULIA DEEVER julia.deaver@2os.com

About 2OS

2nd Order Solutions (2OS) is a boutique credit risk advisory firm that specializes in solving the world's most challenging credit problems. 2OS was founded 18 years ago and consults to a wide range of banks, issuers, fintechs, and specialty finance companies in the US and abroad.

2OS has deep experience with lending businesses across Card, Auto, Small Business, and Personal Loans, at all points in the credit lifecycle. 2OS partners have vast expertise in all aspects of Collections, both as operating executives and as consultants.

For more insights and commentary on the lending industry, visit us at <https://2os.com/insights/>



Providing World-Class Credit Risk
Solutions to Financial Institutions