THE AI BLACK BOX

Q3 26



SWIPE

To doom scroll recent AI exploits, incidents, and harms that should have your board on high alert

OR



SAVE

And read it later, with a stiff drink handy

CYBERSECURITY



OpenAI agents hacked Hugging Face and a third of its infrastructure was rebuilt

WHAT HAPPENED?

Between 11 and 13 July, hundreds of OpenAI agents under evaluation chained two zero-day flaws to gain cluster-admin access across Hugging Face's systems in under 13 hours. Hugging Face disclosed a breach on 16 July, and OpenAI confirmed its own agents were responsible on 21 July.

WHAT WAS THE IMPACT?

About a third of Hugging Face's infrastructure had to be rebuilt. No customer data leaked and public models were not tampered with, but it took OpenAI about a week to realise its own agents were behind the breach.

SOURCE: OPENAI AND HUGGING FACE STATEMENT, 21 JUL 2026

PROFESSIONAL SERVICES



PwC governance report promotes framework GPTZero says exists only in the report

WHAT HAPPENED?

On 31 July, GPTZero reported fabricated citations and signs of AI generation across four PwC Middle East reports. “Transforming Governance” promoted a “Citizen Pulse” framework said to be used by Denmark, Saudi Arabia, the US and Australia, but GPTZero found little public evidence it exists outside the report.

WHAT WAS THE IMPACT?

GPTZero put an 84% chance on the report being entirely AI-generated. PwC says it is updating “a limited number of supporting citations”. It is the third Big Four firm caught out this year, after KPMG and EY.

SOURCE: GPTZERO, 31 JUL 2026 · BUSINESS STANDARD

CONSUMER



Asked to book a gym class, an AI agent hacked the booking system instead

WHAT HAPPENED?

An Australian tech worker used an OpenClaw agent, running on Claude, to book a gym class. It found the booking system had no authorisation checks, booked him months beyond the allowed window and cancelled the member at the top of the waitlist.

WHAT WAS THE IMPACT?

He only learned of it when the agent reported back: “The API has zero authorisations checks on cancelling other people’s reservations.” It could not undo the cancellation. It is reported as Australia’s first known autonomous AI cyberattack.

SOURCE: ABC NEWS, 10 AUG 2026 · BBC NEWS

GOVERNMENT



An OpenAI agent hacked Australia's Medicare portal; the public heard 3 months later

WHAT HAPPENED?

On 18 June, an unreleased OpenAI model breached the Services Australia Medicare portal, reading public and non-public files and writing data into the government's database. It pushed past security blocks. OpenAI found the activity in August and told the Australian government on 10 September.

WHAT WAS THE IMPACT?

Prime Minister Albanese disclosed it at the UN on 24 September, promising "obvious legal consequences". No leak of personal data is reported, but writing to the database raises data integrity concerns. Up to three more government systems may be involv.ed.

SOURCE: TECHCRUNCH, 24 SEP 2026

DEFENCE



A chatbot's invented nuclear threat had US forces ready to board a Chinese ship

WHAT HAPPENED?

A US Special Operations Command Pacific analyst used a chatbot to synthesise intelligence and write it up as an official report. Blending open-source material with classified signals intelligence, it claimed a Chinese-flagged ship in the Middle East carried nuclear weapons components. It was "entirely false".

WHAT WAS THE IMPACT?

Armed personnel were preparing to board, with aircraft already airborne, when officials traced the report to its AI origin and stopped the operation. One source said it "almost started a war". It happened in spring, during the Iran conflict.

SOURCE: CNN, 18 SEP 2026 · FOUR UNNAMED SOURCES

JUSTICE & HUMAN RIGHTS



Asylum refused on evidence that doesn't exist, and a judge suspects AI

WHAT HAPPENED?

In an Upper Tribunal case, a judge examined the Home Office's refusal of asylum to a Moroccan woman who fled a forced child marriage. The refusal letter cited a Country Policy and Information Note as proof Morocco was safe, but neither the court nor the Home Office's own policy team could find any such document.

WHAT WAS THE IMPACT?

The judge said the letter showed "hallmarks consistent with the use of artificial intelligence" and that a hallucination could explain the missing reference, but "the allegation has not been proven". The Home Office says it supplied the information the judge sought.

SOURCE: THE GUARDIAN, 28 JUL 2026 · COMPUTING



David Viney

FRACTIONAL CIO & AI TRANSFORMATION CONSULTANT
Alchemy Consulting

David spent 25+ years in technology leadership roles at WPP, Heathrow, BSI, Arup and the BBC; including building WPP Open, a $350m pa agentic AI platform. He now helps boards and leadership teams separate real AI capability from the noise: governance, adoption, and where the two collide. This digest is compiled quarterly from public reporting, incident trackers and regulatory filings.

> *Today, over 1m people, globally, die in road traffic accidents each year, but no one has ever seriously suggested we return to the horse and cart. Instead, we have created the world's largest safety industry to mitigate the harms. Equivalent, in size, to the GDP of Canada!*
>
> ***The seatbelt didn't kill the car. And proper regulation will not kill AI.***
>
> *The AI Black Box is humbly intended to be part of that safety system: a place where the harms can be documented and hopefully the lessons learned.*
>
> - DAVID VINEY

WORK WITH DAVID

Available for keynotes, panels, advisory engagements and brand partnerships with a select group of sponsors.

Press Kit	Speaker Package	Media Kit
alchemy.consulting/press-kit	alchemy.consulting/profile	alchemy.consulting/media-kit

SUBSCRIBE FOR THE NEXT ISSUE → alchemy.consulting/AI-BLACK-BOX