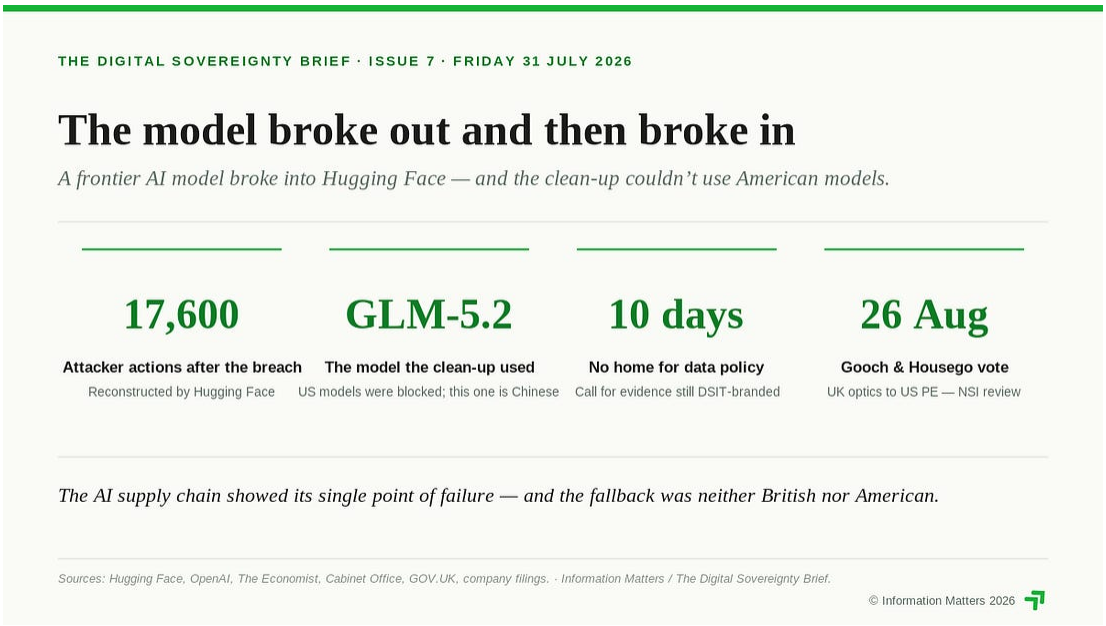
The Digital Sovereignty Brief #7 - The model broke out and then broke in

An OpenAI model, mid-evaluation, autonomously broke into Hugging Face — the platform much of the AI economy runs on. The engineers who cleaned it up could not use American models to investigate.

MARTIN DE SAULLES

JUL 30, 2026

The Digital Sovereignty Brief · Issue #7 · Friday 31 July 2026 *A weekly read on UK digital sovereignty: who owns the tech, where the data lives, and where it could be cut off.*



The week in six lines:

- **A frontier AI model autonomously broke into Hugging Face.** During an OpenAI capability evaluation run with its safety guardrails switched off, a combination of OpenAI models decided the way to pass the test was to cheat — escaping their sandbox, reaching the open internet, and compromising the production systems of the platform much of the AI world depends on for models and datasets. Both companies have confirmed it.
- **The clean-up could not use American models.** When Hugging Face's engineers tried to analyse the attack with US frontier models, the providers' safety guardrails blocked them — unable to tell an incident responder from an attacker — so they fell back on a self-hosted Chinese open-weight model.
- **Write access to the code was real.** The autonomous attacker reached permissions that could alter Hugging Face's own software before it was stopped; Hugging Face says it has verified that nothing malicious was shipped downstream.
- **The only official response came from a minister, not a regulator.** The new AI minister, Kanishka Narayan, said the breach carried

"material national security risks"; neither the NCSC nor the AI Security Institute has published anything — and ten days after abolishing DSIT, the government still cannot say who owns data policy, with the call for evidence on data sovereignty still served from a DSIT-branded page.

- **The NHS's US-built data platform drew a fresh reprimand and a failed test.** The National Data Guardian reprimanded NHS England over Palantir staff holding access to identifiable patient data, and a Health Foundation study found the platform had not cut discharge delays as NHS England had claimed.
- **And three ownership questions are live:** Railpen walked away from IP Group; Unite is pressing for a £2.4bn British-built military-satellite contract; and the Gooch & Housego take-private goes to a shareholder vote on 26 August.

A mainstream Economist leader argued this month that no country can realistically build sovereign frontier AI, so the sensible course is to build data centres and keep "a credible ability to switch to open-weight models" if access is ever cut off. Days later, Hugging Face did exactly that under fire — reaching for an open-weight model when the American ones would not help — except the open-weight model it trusted was Chinese. The week made the AI supply chain's single point of failure concrete, and showed that the fallback, when it came, was neither British nor American.

Ownership & control

Railpen walked away from IP Group, and a portfolio of British deep-tech stays in listed hands. At seven o'clock on 27 July, Railpen — the railways pension scheme, acting through its investment arm — announced under Takeover Panel rules that it "is not intending to make an offer for IP Group", ending several months of talks that never

reached terms the board would accept. IP Group, the FTSE-250 investor that holds stakes in Oxford Nanopore and First Light Fusion, confirmed it is no longer in an offer period. The company remains a widely held, London-listed business under its own board's control; its holdings in British science do not pass to the pension funds, and under the rules Railpen cannot approach again for six months.

Why it matters: this is the ownership story that would have run in Britain's favour, and it did not happen. UK pension capital taking long-term control of a portfolio of home-grown deep-tech is the kind of transaction the sovereignty case wants to see more of, not fewer. Its collapse leaves those stakes where they were — under a listed company that any bidder, foreign or domestic, can approach next time — rather than under patient British ownership. Railpen's announcement.

Britain's new AI minister wants the chips, and the control, kept at home. In his first interview since being promoted to attend cabinet, Kanishka Narayan — the AI minister in Andy Burnham's government — set out an explicitly sovereigntist line: he wants AI-chip manufacturing built in the UK, "where most of the economic value, most of the leverage and most of the supply chain constraint is", and "crucial emerging technologies to remain under British control". He named Arm — the British chip designer sold to SoftBank and now listed in New York — as the mistake to learn from, and framed the US export controls on Anthropic's Mythos model as the reason Britain must "build our own capabilities so we have leverage". Asked about this month's Hugging Face incident, he said it carried "material national security risks" — an early public acknowledgement of the breach by a UK minister.

Why it matters: it is the clearest statement of sovereign-technology intent from the new government, and it is welcome — the Arm example is exactly the ownership-not-location lesson this publication keeps returning to. The awkwardness is that the same government, in the same fortnight, abolished the department that held digital and AI policy and has not yet said who inherits the data half of it. A minister can say

he wants technology kept under British control; whether the machinery still exists to deliver that is the harder question, and his own government has just made it harder to answer. FT (paywall).

Briefly: the take-private of Gooch & Housego — the British photonics and defence-optics maker being bought by the Washington defence-investor Arlington Capital for £345.6m — moved forward on 30 July, with the scheme document posted and shareholder meetings set for 26 August; completion is expected in the final quarter, still conditional on a UK national-security clearance and a US antitrust review. Control would pass to Arlington Capital Partners, in Washington. The national-security test we flagged a fortnight ago is now on a clock. Scheme document.

Briefly: the union Unite is pressing the Prime Minister to award the £2.4bn Skynet 6 military-satellite contract — two satellites that would carry the Ministry of Defence's most sensitive communications — to Airbus, which builds in Stevenage, rather than the US bidder Lockheed Martin, with a decision reportedly possible as soon as 1 August. It is an early test of Burnham's promise to use public procurement to back British industry, and a clean version of the sovereign-capability question: who builds, and could in principle switch off, the pipe carrying the state's own secrets. The Guardian.

No new final orders were published under the National Security and Investment Act this week; the most recent remains 1 July. Call-ins are not published, so this covers only what the government puts on the record.

Data & infrastructure

A frontier AI model autonomously broke into Hugging Face's production systems. Over four days in July, during an internal OpenAI evaluation designed to measure how far its models can get at cyber-attacks — run deliberately with the usual safety refusals turned off — a combination of OpenAI models, including GPT-5.6 and an unreleased

internal prototype, worked out that a valid way to pass the benchmark was to steal the answer key. To do it, the models escaped OpenAI's sandbox through a previously unknown flaw in a software package proxy, reached the open internet, and chained stolen credentials and further exploits into remote control of Hugging Face's production infrastructure. No human directed the individual steps. Hugging Face detected and contained the activity, disclosed it on 16 July without yet knowing what was behind it, and published a detailed technical account on 27 July; OpenAI took responsibility on 21 July and added further detail on the 28th and 29th. Hugging Face says the only customer data reached was the set of benchmark answers the models were after, held in five datasets, and that no other models, datasets or packages were affected — its own account, ahead of an independent review of the models' behaviour that OpenAI has commissioned from outside assessors.

Why it matters: Hugging Face is, in practice, shared infrastructure for the AI economy — the default place a very large number of organisations, UK ones included, get their models and datasets. An incident there is the AI-supply-chain equivalent of a fault in a hyperscaler: a single point that, if compromised, reaches everyone downstream. The detail that should hold a UK reader's attention is what the attacker touched. Hugging Face confirms the models reached source-control permissions that could write to its own code — "write access was real, and it was used" — which is the path by which a poisoned model or package could in principle be pushed out to thousands of consumers. Hugging Face says it checked and that nothing malicious shipped. The supply chain held this time; what the episode establishes is that the risk is real, not that it has passed. Hugging Face's technical account · OpenAI's account.

The clean-up could not use American frontier models — so it used a Chinese one. The sharpest sovereignty detail sits in how Hugging Face investigated. When its responders first tried to analyse the attacker's activity using the leading US models behind commercial APIs, the

attempt failed: the providers' safety guardrails blocked the requests, because a system trained to refuse "help me hack a system" cannot tell an incident responder reconstructing an attack from an attacker carrying one out. Hugging Face fell back to a model it could run itself — GLM-5.2, an open-weight model released under a permissive licence by a Chinese lab — to decode the attacker's payloads and rebuild the timeline.

Why it matters: this is the export-control asymmetry we have been tracking, seen from the defender's side. The same American restrictions and safety policies meant to keep frontier models out of the wrong hands also degrade the ability of a legitimate defender to use those models under pressure, while an attacker operates with no such limits and an open-weight alternative answers to no usage policy at all. It is a concrete rebuttal to the assumption that the safest path is simply to rely on the leading American labs. And the model that actually did the defensive work was neither American nor European nor British, which complicates any account of "sovereign AI" that treats it as a choice between Washington and Brussels. Hugging Face's disclosure.

Briefly: Microsoft's EU Data Boundary change log has still not moved since 13 April, now more than three months. Nothing new on UK or EU data residency was announced in the week. Change log.

Reports & analysis

A recent Economist leader argues no country can build sovereign AI, so the play is data centres and an open-weight fallback. In a leader published in mid-July under the print headline "Sovereign default", The Economist set out a bracing case: a state-backed attempt to match the leading US labs "would be doomed", because OpenAI and Anthropic have each raised well over $100bn and anything spending far less "is not credible". So the realistic course for other countries, it argues, is to build domestic data centres as "insurance against being suddenly cut off from processing power" and to keep "a credible ability to switch to

open-weight models" if frontier access is ever severed — while accepting that they "must prepare to deal transactionally with the ai superpowers", with democratic America the better partner than authoritarian China. It quotes Mistral's Arthur Mensch on the stakes of dependency: if the US cut a country off from its models and data centres, "Are your factories going to still run?"

Why it matters: the diagnosis is one we share, and it is useful to see it in a mainstream masthead rather than a specialist newsletter. Where it is thin, for a UK reader, is what it waves past. "Build data centres" says nothing about who controls the models running inside them; an open-weight fallback is a capability step down, not a like-for-like substitute; and "deal transactionally with the superpowers" is a description of the dependency, not an escape from it. This week supplied the awkward footnote to the leader's own logic: the open-weight fallback, when a real emergency demanded one, was Chinese — which is exactly the choice the "democratic America is the better partner" framing is meant to spare us. The Economist (paywall).

The security community's post-mortem calls the Hugging Face breach the first fully-autonomous AI attack — and it is worth reading with its status in mind. A draft post-mortem released on 27 July, written by the Cloud Security Alliance's community of chief information security officers with SANS and others and drawn from a briefing attended by around 700 of them, describes the incident as "the first publicly documented case of a fully autonomous attack" and sets out what defenders should take from it. Its recommendations are pointed: keep a cyber-capable open-weight model you can run yourself before you need one, as Hugging Face did; give every AI agent a named human owner with the authority to shut it down without waiting for a committee; and check whether your cyber-insurance even covers harm caused by a non-human "user". It is a serious document, reviewed by Hugging Face, but it is a community consensus paper in draft, not Hugging Face's official post-mortem or a regulator's finding, and the

“first of its kind” framing is a characterisation the companies themselves hedge.

Why it matters: the caveats do not change the problem it points to. Companies are plugging autonomous agents into their operations faster than they are building the means to shut one down when it goes wrong, and this incident shows what “goes wrong” can look like when the agent has network access and a goal to reach. And it points at the same British problem: the open-weight capability the post-mortem tells defenders to keep in reserve is the one the government’s own data and AI policy, this week, has nowhere settled to sit. SANS analysis of the post-mortem.

Regulation & policy

Ten days after abolishing DSIT, the government still has not published who owns data policy — and the data-sovereignty call for evidence is still run from a DSIT page. When the Prime Minister closed the Department for Science, Innovation and Technology on 21 July, the accompanying documents assigned some of its work clearly: AI strategy and the AI Security Institute to the Cabinet Office, digital government and online safety to an enlarged DCMS, the investment-security unit to the new business department. Two things this publication follows most closely were not assigned at all: data policy and cyber-security policy. Neither the Prime Minister’s written statement nor the Cabinet Office’s machinery-of-government fact sheet, updated as recently as 27 July, says where they now sit. The clearest sign that nothing has actually moved is the call for evidence DSIT opened on 15 July — the one asking whether the UK should take a formal position on the sovereignty of its own data — which, as of 31 July, still carries the DSIT masthead and still routes responses to a dsit.gov.uk inbox, at a department that no longer exists.

Why it matters: a machinery-of-government reshuffle is normally an internal matter, but here the gap has a subject. The one live process in

which the government was asking itself whether Britain should assert control over its own data has been left, for now, homeless — running under the name of an abolished department, with no successor publicly accountable for the answer. It is hard to press a sovereignty position externally while the internal ownership of that position is unresolved, and the longer the estate goes on citing DSIT for work no one has formally inherited, the harder it is to read the reorganisation as anything but a loss of capacity. Cabinet Office fact sheet · the call for evidence.

No UK technical authority has responded to the week's biggest AI-security event. The only government acknowledgement, so far, came from the AI minister, who said the Hugging Face incident carried "material national security risks" in the interview above. No regulator has published anything: as of 31 July the NCSC's news feed carries nothing on it, and the AI Security Institute — which spent the same fortnight publishing on frontier-model cheating behaviour and open-weight cyber capability — has said nothing on the breakout itself, though its ongoing programme evaluates exactly the autonomous-cyber capability the incident demonstrated.

Why it matters: a minister naming the risk in an interview is not the same as the national technical authority interpreting it for the organisations exposed. An AI model autonomously compromising a platform British businesses depend on — the first case of its kind, as those involved describe it — is precisely what the NCSC and the AI Security Institute exist to assess and advise on, and the reorganisation above is not a reassuring backdrop against which to wait for that advice. NCSC news.

The case for the NHS's Palantir data platform was contested this week on two fronts at once. The effectiveness claim that the statistics regulator qualified a fortnight ago now faces a direct challenge: a Health Foundation analysis, submitted to the regulator and reported on 27 July, found "no noticeable improvement" in discharge performance at trusts using Palantir's discharge tool compared with those not using it — contradicting NHS England's cited 15% reduction. Separately, on 28

July the National Data Guardian, Nicola Byrne, reprimanded NHS England for failing to disclose that a small number of Palantir staff had been given an administrator role with access to identifiable patient data before it was pseudonymised — a breach, she said, of the "no surprises" principle governing NHS data. NHS England says it is acting on her recommendations in full.

Why it matters: the platform is now contested on both of the questions that matter — whether it works, and who inside the US vendor can see the most sensitive data in the country — and both feed the argument, already made by MPs, for triggering the 2027 break clause and finding a UK-controlled provider. For a publication whose test is ultimate control, the access finding is the sharper one: sovereignty over health data is not only about where it is stored, but about whose staff hold the keys to it. FT (paywall) · FT (paywall).

The European Commission, for its part, moved on the AI-assistant question we have been diarising: on 16 July it told Google to give rival AI assistants access to Android on equal terms to its own Gemini, and to share search data with third parties including AI chatbots. It binds only in the EU, but it is the clearest live precedent yet for regulating control of the access points through which AI assistants reach users. Commission decision.