

INFRASTRUCTURE OUTSIDE THE PERIMETER · PART 1 OF 4

The Model You Cannot Audit

AI Dependency and the Regulated Firm

David Newns, Substrate Advisory
July 2026

1. The structural fact

As at the November 2024 supervisory survey, a third of all AI use cases in UK financial services ran on third-party providers, double the share recorded two years earlier, and the top three model providers account for 44 per cent of all named providers across those use cases.¹ The concentration that regulators spent a decade learning to supervise in cloud is forming again, one layer up, in foundation models. A foundation model is a single large model, trained by its provider on broad data, that customers adapt to many different tasks through an interface. As at July 2026, no model provider sits inside any European or UK supervisory oversight regime.

The thesis of this series is that each of four technologies, AI, cloud, public blockchain infrastructure and post-quantum cryptography, is driving changes in the regulatory environment, and that a licensed firm has three postures available: respond, exploit or ignore. For AI the case for responding this year rests on a documented incident. On 12th June 2026 the US Commerce Department, citing national security authorities, directed the suspension of two frontier models, and every firm worldwide that had built on them lost access on a Friday evening with no warning and no route of redress.² Access was restored to general availability nineteen days later through direct negotiation between the vendor and the US executive.³ Model availability is now an operational risk category in its own right, and the firms that had architected for it kept working.

2. The dependency as it stands

The best public data on AI adoption in a regulated financial sector remains the Bank of England and FCA survey of 118 UK firms, published in November 2024, with a further wave underway at the time of writing.^{1,4} Its findings will be broadly representative of any advanced European financial centre. Three quarters of firms were using AI, up from 58 per cent two years earlier. Most adoption is mundane: 62 per cent of use cases are rated low materiality by the firms themselves. The tail is what matters. Sixteen per cent of use cases are rated high materiality, concentrated in risk and compliance, retail banking and general insurance, and over half of all use cases involve some degree of automated decision-making. Foundation models already account for 17 per cent of use cases. For governance, the figure that matters is that 46 per cent of firms reported only a partial understanding of the AI technologies they use, a figure the survey attributes largely to third-party models.¹

The third-party model layer has a precise structure. The weights of the leading closed models, the trained parameters that constitute the model itself, are not public. Access is through an interface. The deploying firm has no sight of the training data, no notice of when the model is retrained, and no ability to pin a version indefinitely. Contractual provisions of the kind the EU's Digital Operational Resilience Act compels reach the vendor relationship but not the model itself. A firm can audit its prompt logs and its outcomes; it cannot audit the thing producing them, and because a provider can retrain the model without notice, an outcome baseline built last quarter may be measuring a different model this quarter.

Open-weight models qualify this picture without dissolving it. A firm that self-hosts an open-weight model can inspect what it runs, keep the workload inside its own jurisdiction and its own data centres, and retrain or pin versions on its own schedule. That deployment pattern is no longer theoretical in banking. BNP Paribas operates a group-wide LLM as a Service platform, hosted in its own data centres on its own GPU capacity, serving open-source models alongside models from Mistral AI, with production use cases in internal assistants, document generation and retrieval.⁵ HSBC announced a multi-year partnership with Mistral AI in December 2025 built explicitly on self-hosted models running on the bank's internal systems.⁶

In my experience operating a licensed financial market infrastructure, the assurance argument for this pattern is genuine: what you host, you can inspect, and what you can inspect, you can evidence to a supervisor.

Two facts stop the open-weight route from being the whole answer. First, open-weight models trail the frontier in capability, and capability is what firms are buying for their most valuable use cases. HSBC's own arrangements illustrate the resulting architecture: the self-hosted European model partnership sits alongside an expanded partnership on agentic AI, systems that take multi-step actions rather than answer single prompts, with a US hyperscaler announced in June 2026.⁷ The observed pattern across the sector is closed frontier models for capability plus self-hosted open-weight models for data-sensitive and continuity-critical workloads, a hedge rather than a substitution. Second, self-hosting moves the operational and assurance burden onto the firm: serving infrastructure, evaluation, patching, security and skills all become internal obligations, and for most firms below the largest tier that burden is decisive.

A further mitigation is increasingly visible. A class of orchestration or gateway components, often called AI gateways or model abstraction layers, sits between the firm's applications and the models. They expose a uniform interface, route requests across providers, log prompts and responses, enforce data-handling policy and in principle allow a model to be substituted by reconfiguration rather than rebuild. The interoperability-standards literature frames this layer as the point at which a firm, or a jurisdiction, can preserve agency over models it cannot inspect.⁸ The argument has substance: a gateway delivers call-level audit data that contractual provisions alone do not secure, and it lowers the cost of moving between providers. Two structural caveats follow. The gateway relocates concentration rather than removing it, since the gateway vendor, the evaluation stack and the underlying compute all remain, and a gateway running on US compute from a US-headquartered vendor leaves any US jurisdictional exposure intact at the layer that has become operationally central. And production switching cost is materially higher than the configuration-change framing suggests: prompts, evaluation suites, agentic scaffolding and safety controls calibrated to one model do not transfer cleanly to another.⁹ The June 2026 suspension tested exactly this machinery, and the firms positioned to come through cleanly were those with a tested fallback chain and, for workloads that could not tolerate any interruption, a self-hosted floor.¹⁰

3. The regulatory response in motion

The regulatory environment around this dependency is moving on four fronts, and a board should be able to place its firm against each.

European Union. ESMA's public statement of May 2024 set the floor that still governs: management bodies remain responsible for decisions taken with the assistance of AI, without exception.¹¹ The EU AI Act has been in force since August 2024.¹² Its trajectory changed materially in the first half of 2026. The Digital Omnibus on AI, proposed by the Commission in November 2025 and adopted by the Parliament on 16th June 2026 and the Council on 29th June 2026, defers the high-risk obligations for stand-alone Annex III systems, which include creditworthiness assessment and credit scoring, from 2nd August 2026 to 2nd December 2027, and for AI embedded in Annex I regulated products to 2nd August 2028. Article 50 transparency obligations still apply from 2nd August 2026, new prohibitions apply from 2nd December 2026, and the AI Office has gained expanded supervisory powers.¹³ The deferral extends the timetable and leaves the substance intact: a bank deploying a foundation model for creditworthiness assessment will carry the Article 26 deployer duties, may become a provider under Article 25 by putting a general-purpose model to a high-risk use under its own name, and will then carry the Article 9 to 15 obligations for risk management, data governance, transparency, accuracy and cybersecurity.¹² Sixteen additional months is

enough time to do the inventory, classification and oversight-evidence work properly, and it is also the period in which firms that treat the deferral as a pause will fall behind.

Alongside the AI Act sits DORA. On 18th November 2025 the European Supervisory Authorities designated the first nineteen Critical ICT Third-Party Providers, bringing the major cloud and platform providers under direct EU-level oversight, with joint examination teams standing up through 2026.¹⁴ No foundation-model provider is among the nineteen. The oversight regime that cloud has entered does not yet reach the model layer, which means the model dependency is governed today only through the deploying firm's own contractual and risk-management obligations.

United Kingdom. The UK has resisted an AI-specific rulebook and instead runs the dependency through existing instruments. The PRA's model risk management principles, SS1/23, in force for banks since May 2024, are technology-agnostic, deliberately cover AI, and make the model inventory, independent validation and senior management accountability the load-bearing controls.¹⁵ The PRA has made AI adoption a 2026 supervisory priority and has stated it will build on SS1/23 during 2026, and the Treasury Committee has recommended that HM Treasury designate the major AI and cloud providers as Critical Third Parties by the end of 2026, with the Treasury stating that initial designation decisions are expected this year.¹⁶ The UK government's Financial Services AI Adoption Plan, published in July 2026, frames domestic AI capability as an answer to sovereign risk and a route to a more diverse and resilient AI ecosystem.⁴ The direction on both fronts is the same: extend the reach of supervision toward the providers, and encourage diversification beneath them.

Switzerland and Germany. FINMA's guidance of December 2024 requires clearly assigned responsibility for AI applications and identifies the explainability of complex models as a central concern.¹⁷ BaFin's guidance of December 2025 places AI within DORA's ICT risk-management framework, which is the clearest statement yet by a European national authority that the AI dependency is to be governed as an operational-resilience matter.¹⁸ The Luxembourg authorities' joint thematic review of 461 institutions, published in May 2025, is the largest disclosed supervisory data-gathering exercise on AI in a European financial sector to date, an example of a supervisor building the adoption picture before writing rules to it.¹⁹

United States. Federal policy runs on a voluntary, industry-collaborative model. Executive Order 14409 of 2nd June 2026 directs agencies to work with the private sector on AI security and continues the approach of voluntary benchmarks and public-private partnership rather than binding pre-release requirements.²⁰ Ten days after it was signed, the same executive branch exercised a different instrument: the export-control directive of 12th June 2026 suspended two deployed commercial models for all users worldwide, and the restoration came through negotiation between the vendor and the government rather than any administrative process a foreign firm could invoke.^{2,3} The same month, a second frontier provider delayed the public launch of its newest model at the US government's request and released it first to a small set of partners whose participation was shared with the government.²¹ Whether any individual action is justified is not the point for a licensed firm outside the United States; a justified withdrawal and an arbitrary one produce identical operational consequences. The structural observation is that access to frontier capability is now mediated by a government to which neither the firm nor its supervisor has standing, and that this mediation has been exercised twice in one month.

4. What response looks like

The response divides into what regulation already compels and what a prudent board does beyond it.

Compelled. Under SS1/23 in the UK, and as a matter of supervisory expectation across the European authorities cited above, every AI system whose output informs business decisions belongs in the model inventory with a materiality classification, an accountable senior manager, independent validation proportionate to materiality and ongoing performance monitoring.¹⁵ Under DORA, the contract with any model provider supporting a critical or important function must carry the Article 30 provisions on audit and access rights, exit assistance, service levels and termination.¹⁴ Under the AI Act as amended, deployers of Annex III high-risk systems must be ready by 2nd December 2027 to operate the system in line with the provider's instructions, assign human oversight to competent persons aware of the risk of automation bias and retain logs; firms whose use of a general-purpose model makes them a provider under Article 25 must be ready to carry the full provider obligations.^{12,13}

Prudent beyond compliance. Six items distinguish the firms that took the June 2026 lesson.

First, a tested substitutability position for every high-materiality model workload: a fallback model actually exercised against the firm's own evaluation suite, with the switching time measured, rather than a clause asserting portability. Second, a firm-owned evaluation suite as a standing asset, because evaluations are what make a model swap a decision rather than a leap, and they are also the only instrument by which a firm can detect that a provider's silent retraining has changed behaviour. The suite does not need to be large to be decisive; it needs to reflect the firm's actual workloads, not public benchmarks. Third, an availability floor for workloads that cannot tolerate interruption: a self-hosted open-weight model, sized honestly against the workload, accepted as less capable and kept warm. The floor is narrow by design: it covers the workloads that cannot stop, not the estate, and it can sit with an in-jurisdiction hosting provider rather than on the firm's own premises, which is what keeps it within reach of firms below the largest tier. Fourth, contractual terms at the model boundary that go beyond the DORA minimum: notice of material model updates, version pinning windows, retention and data-use terms compatible with the firm's own obligations, and express treatment of government-directed suspension as a vendor-side event with defined consequences. The June 2026 suspension supplies the drafting precedent: a government-directed withdrawal is no longer a theoretical clause, and a vendor unwilling to engage on it is telling the firm where the risk sits. Fifth, human-oversight evidence designed to survive inspection: named individuals, documented competence, recorded interventions and override rates reported upward, because the AI Act states the oversight requirement but supplies no verification method, and the firm that can evidence exercised oversight rather than recorded oversight will set the standard the rest are measured against. A register of named overseers exists in most firms already; a record of interventions actually made, with override rates someone has examined, is rarer, and it is what an inspection team will ask for. Sixth, board-level reporting: a quarterly view of model concentration by provider, high-materiality use cases, evaluation coverage and tested-fallback status, in the same pack as the cloud concentration numbers. Placing them together is deliberate: concentration one layer up is the same risk category, and a board that sees the two side by side will price them together.

None of this requires believing that closed frontier models should not be used. It requires treating the model layer with the same discipline the sector already applies to cloud, before a designation regime compels it.

5. The advantage case

The advantage evidence for early movers is real but thinner than the equivalent case for cloud, and it should be stated at its actual weight.

The strongest element is continuity. The June 2026 suspension tested the logic in production conditions: open-weight fallback candidates were available within days of the withdrawal, and the architecture positioned to use them was the one with a tested fallback chain and a self-hosted floor already in place.¹⁰ Firm-level outcome data from the window is not yet published, and the continuity case should be carried at that weight. The second element is supervisory positioning. With the PRA naming AI a 2026 supervisory priority and CTP designations expected in the UK this year, firms able to produce a complete model inventory, evaluation coverage and tested substitutability in their next supervisory dialogue will spend that dialogue on their business rather than on remediation, and the same artefacts serve licensing applications and counterparty due diligence, where AI governance questionnaires are now standard. The third element is the sixteen-month AI Act runway: conformity work done against a settled deadline is cheaper than conformity work done in a compressed scramble, and grandfathering provisions reward systems placed on the market before the new deadlines.¹³ What I cannot evidence is a funding-cost or revenue premium attaching to AI governance maturity as at July 2026; if that premium emerges it will show first in due-diligence outcomes, and boards should not build the business case on it yet.

6. The cost of ignoring it

The documented record is recent and sufficient.

The June 2026 suspension is the anchor case: a frontier model withdrawn from every customer worldwide three days after launch, restored to general availability nineteen days later, with no supervisory channel involved at any point.^{2,3,22} Any firm that had moved a material workflow onto that model in its launch week carried an unplanned outage of its own making, in the sense that the single-provider architecture was a choice. The October 2025 US-EAST-1 cloud outage sits one layer down as the reminder that the infrastructure beneath the models carries the same concentration: a single-region failure whose reported UK impact included a major bank and the tax authority.^{23,24} The outage drew parliamentary questions to HM Treasury on why UK banking and tax infrastructure had been interrupted from a US region, and on whether the provider should be designated a Critical Third Party, which is the concentration argument made for the firm by its own legislature.²⁴ And the supervisory finding that should concentrate a board's attention costs nothing to look up: nearly half of surveyed firms report only partial understanding of the AI they already run.¹ There is, as at July 2026, no published European enforcement action specifically for AI governance failure at a financial institution; on the current trajectory of supervisory priority-setting, designation regimes and the AI Act's staged application, the first such action is a matter of timing, and the firms that appear in it will be the ones whose inventories, oversight evidence and exit positions were paper exercises.

7. Board questions

Six questions a board member should be asking the executive now.

1. Do we have a complete inventory of AI systems whose output informs business decisions, with materiality ratings, and when was it last reconciled against what the businesses actually run?
2. For each high-materiality use case on a third-party model, have we tested a fallback against our own evaluation suite, and what was the measured switching time?
3. Which of our AI uses would make us a provider rather than a deployer under Article 25 of the AI Act, and what is the plan to carry those obligations by 2nd December 2027?

4. If our primary model provider suspended service tonight, by government direction or otherwise, which workflows stop, and what is our floor?
5. What notice do our model contracts give us of retraining or material model changes, and how would we detect an unnotified change in behaviour?
6. What does the board see, and how often, on model-provider concentration, and is it in the same pack as our cloud concentration reporting?

8. Sources and notes

1. Bank of England and FCA, "Artificial Intelligence in UK Financial Services 2024," 21st November 2024 (118 respondents; top three providers account for 73 per cent of cloud, 44 per cent of model and 33 per cent of data providers across all AI use cases; provider identities anonymised): <https://www.bankofengland.co.uk/report/2024/artificial-intelligence-in-uk-financial-services-2024>
2. Anthropic, "Statement on the US government directive to suspend access to Fable 5 and Mythos 5," 12th June 2026: <https://www.anthropic.com/news/fable-mythos-access>
3. Anthropic, "Redeploying Claude Fable 5," 30th June 2026 (export controls lifted 30th June 2026; general availability restored 1st July 2026; Mythos 5 restored from 26th June 2026 to a vetted set of US organisations, with wider access under discussion): <https://www.anthropic.com/news/redeploying-fable-5>
4. HM Treasury, "Financial Services AI Adoption Plan," July 2026 (notes that further survey work by the FCA and Bank of England is underway to update the 2024 adoption figures; frames domestic AI capability as addressing sovereign risk and supporting a more diverse and resilient AI ecosystem): <https://www.gov.uk/government/publications/ai-adoption-plan-financial-services/financial-services-ai-adoption-plan>
5. BNP Paribas, "BNP Paribas provides its businesses with an LLM as a Service platform," June 2025 (hosted in the group's data centres with dedicated GPU capacity; open-source models alongside Mistral AI models; internal assistants, document generation and document retrieval in production or experimentation): <https://group.bnpparibas/en/press-release/bnp-paribas-provides-its-businesses-with-an-llm-as-a-service-platform-to-accelerate-the-industrialization-of-generative-ai-use-cases>
6. HSBC, "HSBC and Mistral AI join forces," 1st December 2025 (multi-year partnership; self-hosted AI models operating on HSBC's internal technology systems): <https://www.hsbc.com/news-and-views/news/media-releases/2025/hsbc-and-mistral-ai-join-forces-to-accelerate-ai-adoption-across-global-bank>
7. The Asian Banker, "HSBC expands Google Cloud AI partnership," June 2026 (expanded agentic AI partnership on Google Cloud announced June 2026, alongside the December 2025 self-hosted Mistral arrangement): <https://www.theasianbanker.com/updates-and-articles/hsbc-expands-google-cloud-ai-partnership-with-100-million-use-case-threshold>
8. Donahoe E and Komaitis K, "Why AI Sovereignty Depends on Interoperability Standards," Tech Policy Press, 17th February 2026: <https://www.techpolicy.press/why-ai-sovereignty-depends-on-interoperability-standards/>
9. The Future Society, "Beware of Geeks Bearing Gifts: Building True EU Frontier AI Sovereignty," April 2026, characterises sovereignty offerings that simulate autonomy while deepening dependency: <https://thefuturesociety.org/eu-frontier-ai-sovereignty-report/>
10. The New Stack, "Fable 5 ban: 4 open models responded before Anthropic could restore access," June 2026 (open-weight releases as fallback candidates during the suspension window, including an Apache 2.0 licensed model positioned for self-hosting on a single GPU): <https://thenewstack.io/fable-ban-open-weights/>
11. ESMA public statement on AI in MiFID II investment services, ESMA35-335435667-5924, 30th May 2024: <https://www.esma.europa.eu/document/public-statement-ai-and-investment-services>
12. EU AI Act, Regulation (EU) 2024/1689, in force 1st August 2024: <https://eur-lex.europa.eu/eli/reg/2024/1689/oj>
13. Freshfields, "EU AI Act unpacked #34: The final Digital Omnibus on AI," July 2026 (Parliament adoption 16th June 2026, Council adoption 29th June 2026; Annex III high-risk obligations deferred to 2nd December 2027, Annex I to 2nd August 2028; Article 50 transparency from 2nd August 2026; new prohibitions from 2nd December 2026; expanded AI Office powers): <https://www.freshfields.com/en/our-thinking/blogs/technology-quotient/eu-ai-act-unpacked-34-the-final-digital-omnibus-on-ai-key-amendments-to-the-a-102nber> Corroborated by Gibson Dunn, "EU AI Act Omnibus Agreement," May 2026: <https://www.gibsondunn.com/eu-ai-act-omnibus-agreement-postponed-high-risk-deadlines-and-other-key-changes/>

Council of the EU press release, final approval, 29th June 2026: <https://www.consilium.europa.eu/en/press/press-releases/2026/06/29/artificial-intelligence-council-gives-final-green-light-to-simplify-and-streamline-rules/> Final act signed 8th July 2026; awaiting Official Journal publication as at 18th July 2026. This entry must be replaced with the Official Journal citation on publication.

14. Joint ESAs communication, designation of nineteen Critical ICT Third-Party Providers under DORA Article 31, 18th November 2025: <https://www.eba.europa.eu/publications-and-media/press-releases/european-supervisory-authorities-designate-critical-ict-third-party-providers-under-digital>
15. Prudential Regulation Authority, SS1/23 "Model risk management principles for banks," published 17th May 2023, in force 17th May 2024: <https://www.bankofengland.co.uk/prudential-regulation/publication/2023/may/model-risk-management-principles-for-banks-ss>
16. House of Commons Treasury Committee, "AI in financial services: Responses to the Committee's First report," 2026 (Committee recommendation to designate major AI and cloud providers as Critical Third Parties by end 2026; HM Treasury expects initial designation decisions this year; PRA 2026 supervisory priorities include AI adoption and the PRA intends to build on SS1/23 in 2026): <https://publications.parliament.uk/pa/cm5901/cmselect/cmtreasy/1791/report.html>
17. FINMA Aufsichtsmitteilung 08/2024, 18th December 2024: <https://www.finma.ch/en/news/2024/12/20241218-mm-finma-am-08-24/>
18. BaFin, "Guidance on ICT Risks in the Use of Artificial Intelligence at Financial Entities," 18th December 2025: https://www.bafin.de/SharedDocs/Veroeffentlichungen/EN/Meldung/2025/meldung_2025_12_18_orientierungshilfe_ikt_risiken_en.html
19. BCL and CSSF joint thematic review on AI in the Luxembourg financial sector, May 2025 (461 respondent institutions, an 86 per cent response rate; fieldwork June to August 2024): <https://www.cssf.lu/en/2025/05/second-thematic-review-on-the-use-of-artificial-intelligence-in-the-luxembourg-financial-sector/>
20. Executive Order 14409, "Promoting Advanced Artificial Intelligence Innovation and Security," 2nd June 2026, Federal Register Vol 91 No 108, 5th June 2026, FR Doc 2026-11415: <https://www.govinfo.gov/content/pkg/FR-2026-06-05/pdf/2026-11415.pdf> Congressional Research Service explainer: <https://www.congress.gov/crs-product/IF13268>
21. OpenAI, "Previewing GPT-5.6 Sol: a next-generation model," 26th June 2026 (limited preview at the US government's request for a small group of trusted partners whose participation was shared with the government, ahead of broader availability): <https://openai.com/index/previewing-gpt-5-6-sol/>
22. Anthropic, "Claude Fable 5 and Claude Mythos 5," 9th June 2026 (launch announcement): <https://www.anthropic.com/news/claude-fable-5-mythos-5>
23. AWS post-event summary, US-EAST-1 service disruption, 20th October 2025 (latent race condition in DynamoDB automated DNS management): <https://aws.amazon.com/message/101925/>
24. Computer Weekly, "Government faces questions about why US AWS outage disrupted UK tax office and banking firms," October 2025 (UK impact including HMRC and Lloyds Banking Group; parliamentary questions to HM Treasury including on Critical Third Party designation of the provider): <https://www.computerweekly.com/news/366633473/Government-faces-questions-about-why-US-AWS-outage-disrupted-UK-tax-office-and-banking-firms>

David Newns is an Advisor to BPX Markets Limited and an Independent Non-Executive Director of Axiology Group. The views expressed are personal and do not represent the positions of any organisation with which he is associated. Nothing in this report constitutes legal, regulatory or investment advice.