

By



Colin Payne

Head of Innovation at the Financial Conduct Authority (FCA) and Chair of the Global Financial Innovation Network (GFIN)

Colin Payne is a distinguished leader in financial innovation, currently serving as the Head of Innovation at the Financial Conduct Authority (FCA) and Chair of the Global Financial Innovation Network (GFIN) convening over 90 national regulators to build a powerful, collaborative network. In these roles, he spearheads strategic initiatives to foster growth and enhance global competitiveness in financial services. Before joining the FCA, Colin was the Senior Director for Central Banking and Fintech Product Lead at PwC, where he delivered strategic innovation programs across the Gulf Cooperation Council, including national payment strategies and digital asset initiatives. With over 25 years at the forefront of technology and change, Colin has a proven track record of designing and deploying world-class digital banking solutions for global brands such as Lloyds Banking Group, HSBC, Barclays, and Wells Fargo. Colin's dynamic leadership and open style inspire teams to push boundaries and unlock potential, shaping the future of financial services in the UK and beyond. He is a visiting lecturer at King's College and holds an honorary position at Middlesex University.

Explore more from
The Kyndryl Institute
kyndryl.com/institute

The trust inversion

The year the infrastructure arrived: The forecast, corrected

A year on, the claim I could not yet prove looks less like a forecast and more like a description, though not quite in the way I expected.

Last year, in an essay for The Kyndryl Institute, I closed with a line I believed but could not yet prove: the agentic era of financial services was here, and the time to act was now.

In September 2025 I described a “Do It For Me” economy in the future tense: consumers delegating to AI proxies that shop, switch and settle on their behalf. Less than a year later, that economy no longer looks theoretical. By the spring of 2026, agent-initiated commerce had moved from demonstration into pilots, beta launches and early production integrations. Visa and Mastercard were building dedicated agentic payment infrastructure: tokenisation, authentication and spending controls designed for a machine acting on a customer's behalf. OpenAI had introduced checkout inside ChatGPT, while keeping acceptance, payment processing, fulfilment and customer support with merchant backends. Amazon was testing a feature that completes purchases from other brands' sites on a customer's instruction. The networks' own language now treats checkout as something agents will increasingly initiate, not merely assist.

Moving into pilots, beta & early production — 2025-26

Visa Intelligent Commerce

Mastercard Agent Pay

OpenAI in-chat checkout

Amazon Buy for Me (beta)

It would be easy to file this under retail payments. That would be a mistake. The same architecture that lets an agent buy a suitcase can let it move money, switch products, triage a complaint, manage a credit line, execute a trade or assemble compliance evidence. Commerce is simply where the rails were laid first, because that is where the demand and the revenue were clearest. Financial services is where the consequences land hardest.

So the headline I might have written a year ago (autonomy is coming) needs one important correction. Full autonomy has not arrived everywhere; across most of finance it has barely begun. What arrived is the infrastructure for autonomy: identity, payment authority, tool access, the ability to initiate a transaction. And it arrived before the agents using it had earned broad trust. That is the trust inversion: the rails of delegation are being built faster than the reliability of the things we are delegating to.

The inversion: Dependability is now the constraint

Here is what I did not see clearly enough. In 2025 the implicit worry (mine included) was capability. We pictured agents that would be too good and too fast: superhuman traders moving in microseconds, optimization we could not follow. The risk we braced for was machines outrunning us.

The risk that actually showed up is more mundane and more awkward. The agents are not too capable. They are not yet dependable enough.

It is worth separating two things that engineers never confuse. **Capability** is what a system can do on a good day. **Dependability** is what it does on every other day: whether it fails rarely, fails safely, and tells you when it has failed. We do not certify an aircraft on its best flight; we certify it on its worst plausible one. Agentic AI is being deployed on something closer to the opposite basis: impressive demonstrations, thin evidence of behavior under stress.

The evidence on dependability is sobering. On the most demanding independent benchmark of real office work, the strongest agent completes around a third of tasks autonomously (and only at the top of the range, depending on how partial credit is scored). MIT's widely-cited study found that roughly 95% of organizations were seeing no measurable return from their generative-AI deployments; a finding about adoption broadly, but a caution that bites hardest on anything we let act unsupervised. Gartner expects more than two in five agentic-AI projects to be cancelled by the end of 2027, defeated by cost, unclear value or inadequate controls.

The reliability gap, by the numbers

≈30% of real office tasks the strongest agent completes autonomously (higher with partial credit) Carnegie Mellon · TheAgentCompany, 2025

95% of organisations report no measurable return from generative-AI deployments
MIT · Project NANDA, 2025

40%+ of agentic-AI projects expected to be cancelled by end-2027 Gartner, 2025

These are not the numbers of a workforce ready for unsupervised delegation. They are the numbers of a talented graduate who cannot yet be left alone with the client. And yet the rails got built anyway. That is the inversion stated plainly: we have constructed the infrastructure of delegation for a generation of agents whose binding limitation is reliability. **The danger in 2026 is not the brilliant agent acting too fast. It is the mediocre agent, deployed too widely, trusted too soon**, in a system wired so that its mistakes do not stay local.

Because the failure mode of an agent is not a single bad decision. It is propagation. An error does not sit still: it flows downstream through every system the agent can touch, and into every other agent that consumes its output. A wrong human decision is a loss. A wrong agentic decision, plugged into live rails, is a contagion vector.

The systemic concern follows directly, and it is not exotic. It is correlation. If thousands of firms build their agents on a handful of foundation models, served by a handful of cloud providers, they do not hold independent risks; they hold the same risk many times over. One flawed assumption, one poisoned data source, one model update can express itself everywhere at once. Correlation, not individual error, is the transmission mechanism: the lesson 2008 taught about correlated exposures, now rehearsed in software. This is precisely why the Bank of England has begun simulating how AI agents might “herd” under stress, and why bodies from the Financial Stability Board to the Bank for International Settlements have moved concentration and correlated behavior to the center of their work on AI.

The architecture of delegation: Identity is necessary; permission is the point

Strip away the novelty and the problem is an old one. Economists call it the principal-agent problem: how does a principal get an agent to act in the principal's interest, when the agent has its own latitude and the principal cannot watch every move?

For a century, the discipline of delegation has rested on three things; human finance encodes them in contracts, fiduciary duty, supervision and, ultimately, the courts.

The delegation problem

Mandate: bounds what the agent may do

Monitoring: reveals what it actually did

Recourse: acts when it works against you

Software agents arrive with none of the three by default. The architecture must now supply all of them.

Software agents arrive with no licence, no fiduciary duty, no fear of being struck off. The entire burden of mandate, monitoring and recourse therefore falls on the technical architecture instead. That is why identity matters, and also why identity, on its own, is not enough.

Identity answers one question: who acted? It is necessary; without it, accountability has nowhere

The constraint on agentic finance is no longer how clever the agents are. It is how far they can be trusted, and trust has become the scarce asset.

to attach. But the question a regulator will press harder is the next one: what were they allowed to do? Identity without permission is a name without a leash. Permissions answer the questions that actually decide whether a delegation was safe, and they are the questions firms should be designing for now.

What "good" looks like: The accountability stack

1. Verifiable mandate

Provable authority to act, traceable to a named, accountable principal.

2. Scoped permissions

Hard limits on value, purpose and duration (not merely a stated intent).

3. Tamper-evident audit

A reconstructable record of what was done, and on what basis.

4. Real-time revocation

Consent that can be narrowed or withdrawn instantly, not at renewal.

5. Mapped liability

A clear answer, decided in advance, to “who pays when this fails?”

6. Redress route

A defined way for a harmed customer to be made whole.

7. Kill switch

The ability to halt the agent, individually and at system scale.

This is why the quiet contest over standards matters more than any model release. The protocol that connects agents to tools and data (Anthropic's Model Context Protocol, donated to the Linux Foundation's Agentic AI Foundation in December 2025) has become one of the foundational pieces of plumbing. But it is only one layer. The harder, more fragmented landscape is forming around payments

and authorisation, where competing schemes are still being fought over: the Trusted Agent Protocol, the Agentic Commerce Protocol, Google's Agent Payments Protocol, the Machine Payments Protocol and the Universal Commerce Protocol. Whoever sets those standards is, in effect, designing the accountability chain: deciding whether every agentic action can be traced back to an authorising human, or whether responsibility dissolves into the gaps between systems. That is a public-interest question being settled, for now, largely by private firms. Regulators need not write the protocols. But we cannot be absent from the room in which they are written.

From forecasting to supervising: The forecast became a workplan

A year ago, the FCA's AI Lab was an instrument for understanding something that had not fully arrived. Its job now is different. We are no longer only forecasting agentic finance; we are beginning to supervise its first live applications.

The shift is visible in our own program. The first cohort of the Supercharged Sandbox (firms building and stress-testing real systems under genuine constraints, not slideware) completed and showcased its work on 28-29 January 2026. AI Live Testing, which lets firms run AI in live markets under supervision, ran its second application window from January to March and began testing from late April. And in the latest cohort in our Supercharged Sandbox we said plainly what we most want to see: agentic systems, among them agentic payments, compliance agents and customer-facing agents. We are deliberately pulling the hardest cases towards us, where we can watch them, rather than waiting to read about them in an enforcement file.

That reflects a change in posture our Chief Executive has signalled: a different relationship between regulator and regulated. The direction of travel is clear: honest failure inside a tested, controlled system should be treated differently from negligent failure left unmanaged. That distinction is,

I think, the most useful one a regulator can draw in this period.

We are not drawing it alone. The Bank of England's Financial Policy Committee has reached a careful but pointed conclusion: advanced AI is not yet adopted in a way that presents systemic risk, but those risks could build quickly; agentic AI in payments and markets is exactly where private incentives may part company with system-wide stability. Parliament's Treasury Committee has gone further on concentration, recommending that the largest AI and cloud providers be designated critical third parties by the end of 2026; the Government has said it is gathering evidence and expects initial decisions this year. The concern I raised last year, about a few providers sitting beneath the whole system, is now a live question of statutory designation.

The UK's actual edge: Proximity is not a thicker rule book

It was never likely that the UK would win this era by writing the longest rulebook. The UK confirmed in 2025 that it would not create bespoke AI legislation, governing instead through the principles already in place: the Consumer Duty, the Senior Managers regime, outcomes over instructions. I defended that choice then and I defend it now. A rulebook written for today's agents would be obsolete before the ink dried, and the agents of 2027 will be different again.

But “no new law” was never the same as “no plan.” Our edge is not regulatory restraint; it is regulatory proximity: being the jurisdiction where agents are tested, in the open, before they are trusted at scale. The year's lesson is that this was the right bet, because the binding problems turned out to be reliability, accountability and controlled delegation, and none of those is solved by statute. They are solved by evidence: by watching agents fail safely, learning what good looks like, and codifying it into expectations firms can actually meet.

We do not reward autonomy. Raw autonomy is becoming cheap. **We** reward agents that can prove they are reliable, and systems that can prove who is accountable when they are not.

We do not reward autonomy.
Raw autonomy is becoming
cheap. We reward agents that
can prove they are reliable,
and systems that can prove
who is accountable when
they are not.

The next phase: Closing the inversion

If the last year was about recognizing the inversion, the next is about closing it. That work is concrete, and it is mostly not about building cleverer models.

1. Build the accountability stack before the volume arrives

Mandate, permissions, audit, revocation, liability and redress should be designed in now, while agentic finance is still in pilots, not retrofitted after the first scandal. The cheapest moment to require a kill switch is before anything needs killing.

2. Treat correlation as a first-class risk

If the system is going to lean on a few models and a few providers, then concentration, common-mode failure and agent herding belong inside stress tests and the critical-third-parties regime, not in a footnote. Diversity of models and suppliers may need to be encouraged as a stability property, rather than left to procurement.

3. Make the standards a public-interest project

The protocols for agent identity, permission and payment are being set this year. The UK should be an active voice in shaping them toward traceable accountability, and should be willing to say plainly that an agent which cannot prove its mandate has no business initiating a regulated transaction.

4. Coordinate across borders, because agents will not respect them

An agent authorized in one jurisdiction will transact in another within the same second. Without interoperable identity and shared supervisory expectations, we will reproduce (at machine speed) every gap and arbitrage that fragmented regulation has always created. Through GFIN and its peers, cross-border coordination on agentic finance is not a nicety. It is infrastructure.

The race was never to build the most autonomous financial system. It is to build the most trustworthy one: fast enough to matter, carefully enough to last.

A year ago I wrote that agentic AI was here and the time to act was now. The infrastructure

for autonomy has indeed arrived, and we did begin to act. The task of the coming year is to make these agents worthy of the trust we are already, quietly, every day, beginning to place in them. —

The race was never to build the most autonomous financial system. It is to build the most trustworthy one: fast enough to matter, carefully enough to last.

References

1. Colin Payne, Agents of change: Agentic AI and the future of financial services, The Kyndryl Institute, 30 Sept. 2025. kyndryl.com/gb/en/institute/agents-of-change
2. Visa, Visa and Partners Complete Secure AI Transactions, Setting the Stage for Mainstream Adoption in 2026, Visa newsroom, Dec. 2025 (Intelligent Commerce; Trusted Agent Protocol). Mastercard, Agent Pay, 2025–26. [visa.com](https://visa.com/mastercard.com) · mastercard.com
3. OpenAI, Instant Checkout in ChatGPT: release notes, 2025 (acceptance, payment processing, fulfilment and customer support retained by merchant backends). Amazon, Buy for Me (beta), 2025–26.
4. Carnegie Mellon University, TheAgentCompany: Benchmarking LLM Agents on Consequential Real-World Tasks (updated 2025); CMU news, June 2025. arxiv.org/abs/2412.14161
5. MIT Project NANDA, The GenAI Divide: State of AI in Business 2025, July 2025.
6. Gartner, Predicts 2025: over 40% of agentic-AI projects to be cancelled by end-2027, 2025.
7. Anthropic / Linux Foundation, Model Context Protocol donated to the Agentic AI Foundation, Dec. 2025. For agentic commerce and payment protocols, see Google's Agent Payments Protocol (AP2), the OpenAI / Stripe Agentic Commerce Protocol, Visa's Trusted Agent Protocol, and related machine-payment and universal-commerce protocol developments, 2025–26.
8. Bank of England, Financial Policy Committee Record, Apr. 2026; S. Breeden, correspondence to the Treasury Committee on AI and financial stability, Apr. 2026.
9. Financial Stability Board, Monitoring Adoption of AI and Related Vulnerabilities in the Financial Sector (Oct. 2025) and The Financial Stability Implications of AI (Nov. 2024).
10. House of Commons Treasury Committee, Artificial intelligence in financial services (report and Government response), 2026.
11. FCA, AI Lab: Supercharged Sandbox and AI Live Testing programme pages and updates, 2025–26. fca.org.uk/firms/innovation/ai-lab

