14 November 2026

Madhuvarshitt Yerramilli Subramanya · GFTN

GFTN

PROVE before you deploy

ALFIN playbook series v5 · November 2026

Prioritise · Risk-assess · Orchestrate · Validate · Embed

Contents

Executive summary

1. Money is uniquely circular

2. Fluent, fast and unevenly reliable

3. When answers become actions

4. PROVE before you deploy

5. Five closed loops: put PROVE to work

6. The deployment decision

Sources

Author, publisher and disclaimer

Five stages before a system may price, block, move or deny money.

Executive summary

In finance, a decision does two things at once. It moves value the moment it is carried out, and it helps shape the evidence later used to judge it. A loan that is refused leaves no repayment record; a payment that is blocked leaves no fraud loss. An AI system given authority over such decisions can act before anyone checks, on evidence it partly wrote itself.

That is the risk. AI is not useless in finance: the newest models are fast and often good, but unevenly so, and they can sound just as sure when they are wrong. In a March 2026 test of fintech support tasks, the best agent completed 25.52% of tasks on a first attempt [s21]. In a credit simulation, a model retrained only on the applicants it approved grew more accurate while the share of defaulters it caught fell from about 0.37 to 0.13 [s7]. A dashboard would have shown improvement.

Most firms are still cautious. In the Bank of England and FCA survey published in October 2026, no more than 9% expected agents to approve or decline customers, or to initiate, execute or block payments, even counting plans for the next 12 months [s22]. The question is how to move further without losing sight of what the system hides.

Our answer is PROVE, five things an institution should show before a system may price, block, move or deny money:

- **Prioritise:** the use is worth it, against a baseline measured before launch.
- **Risk-assess:** it knows what can go wrong in both directions, and which outcomes will become harder to see.
- **Orchestrate:** the limits sit in the systems around the model, not in its instructions.
- **Validate:** the system performs on evidence it did not choose.
- **Embed:** named business, technical, risk and compliance owners answer for the outcome.

PROVE decides how much a system may do, and when that authority is taken back. It cannot tell a lender how the people it never lent to would have repaid; no checklist can, and Section 4 says what to do when that gap cannot be measured. The more a system is allowed to do, the more an institution should be able to show.

Sections 1 to 3 set out the problem and the four steps from answering to acting. Section 4 sets out PROVE. Section 5 applies it to credit, fraud, anti-money laundering, collections and trading, each with the measure that keeps the headline number honest. Section 6 turns it into a decision: a written permission at a stated step, with triggers that narrow it and a date on which it lapses unless renewed.

1. Money is uniquely circular

Most businesses use money to get something else. A baker spends it on flour, sells bread and counts the takings at the end of the day. The Bank of England lists the jobs money does: it is a medium of exchange between buyer and seller [s1], it holds value you can spend later [s1], and it gives everyone a shared standard for prices [s1]. In finance, money does all three jobs and is also the product. A bank lends money, charges for it in money and keeps score in money.

That double role puts two things inside every financial decision. The decision moves value the moment it is carried out, and it helps shape the evidence later used to judge it. Our concern follows from both: an AI system given authority over such decisions can act before anyone checks, on evidence it partly wrote itself.

The ledger is the vault

In a factory a bad decision has to travel. The wrong parts get ordered, a truck is loaded, a pallet turns up at the dock, and somewhere along that chain a person may notice and stop it. In finance the instruction and the value sit much closer together. Change the record and you have changed who controls the money, often without anything leaving a building.

In October 2022 attackers forged proofs on BSC Token Hub, a cross-chain bridge on BNB Chain, and withdrew 2 million BNB [s2], worth about US$570 million at the time [s3]. Nobody carried anything out of a vault. A bug in the bridge's smart contract let them forge transactions and send the tokens to their own wallets [s3]. A regulated bank has payment limits, reconciliation and people on call that a crypto bridge lacks.

What matters is where those safeguards sit. A check that runs before an instruction is executed can stop it; one that runs after the money has moved leaves the institution trying to get it back. That is the gap between an answer and an instruction. A recommendation waits for someone to decide, while a system with authority to act carries the decision out.

Harm in both directions

A credit decision can go wrong in two directions, and both land on a person.

Lending can hurt when the repayments outrun the household. In January 2025, at Tarihal near Belagavi in Karnataka, a microfinance company used a court order to seize a house after six months of missed instalments and put a mother and her newborn out of it [s4]. Her husband had borrowed to build the house, and the lender wanted the remaining Rs 7.5 lakh (Rs 750,000) in one go [s4]. A state minister later helped the family move back in [s4].

Refusing can hurt too, and it leaves less of a trace. In Germany a score from SCHUFA, a credit bureau, led a bank to refuse a woman a loan [s5]. When she asked what data sat behind it, SCHUFA gave her the score, 85.96%, and little else, citing business confidentiality [s5]. Her case reached the EU's top court [s5], which ruled that a score like this can itself count as an automated decision when a lender leans heavily on it [s6].

Neither story says what the right answer was.

> ***Try it: Meena's loan.*** *Meena runs a home bakery and asks for a working-capital loan to buy flour, packaging and an oven part before the festival season, when she makes most of her year's sales. Her file shows 14 months of credit history (two small accounts, a thin file), uneven income (low most months, high at festival time) and a model score of 0.58 against your cut-off of 0.60. Would you approve or decline? Approve, and her repayments become evidence. Decline, and the loan leaves no record of what would have happened. Section 6 comes back to her. An illustration; on the web edition this is an interactive box.*

What the lender gets to see

Approving or rejecting an application also decides what the lender will learn.

Approve a loan and, over the following year, the bank watches the repayments come in or fail to. Reject it, and the loan produces no repayment history at all. Other channels may offer clues about how the applicant fared, but the direct answer is gone. The bank's future evidence comes mostly from the people it already chose to trust.

What a low default rate cannot show

Suppose the approved book performs well. Defaults are low, the dashboard is green and the credit committee is pleased. That is useful evidence about the loans the bank made. It says much less about the applicants it turned away, some of whom would have defaulted and some of whom would have repaid in full and grown. The default rate cannot tell those groups apart, because neither got the loan.

None of this proves the rejections were wrong. Creditworthiness, affordability and the borrower's wider welfare are separate questions, and a clean repayment record answers only part of the first. A borrower can pay every instalment on time and still end up worse off for having borrowed.

The gap can widen when a model is retrained on its own selections. A 2026 simulation study retrained a credit model 10 times, each time only on the applicants it had approved. Its accuracy crept up from about 0.820 to 0.835. Its recall, the share of real defaulters it caught, fell from about 0.37 to 0.13 [s7]. Inside the study, the researchers knew how every refused applicant would have fared, so they could count what the model missed. A real lender has no such record. A second model in the same study saw every applicant's true outcome. It ended with lower accuracy, about 0.816 [s7], yet its recall stayed near where it started [s7]. The first model's training data also understated the default rate: 17.0% against a true 22.1% [s7]. The authors call this "an illusion of improvement" [s7].



Figure 1: The model that saw less looked better. Model A learns only from the loans it approved. Model B sees every applicant's true outcome. [s7]

Filling in the missing outcomes by guesswork did not cure it. One common method, Simple Extrapolation, labels each rejected applicant with the model's own prediction. In the study it reversed the distortion instead of removing it: by the paper's own summary measure, where training only on approved loans understated the default rate by 22%, Extrapolation overstated it by a symmetric 22% [s7]. Worse, a team choosing by the usual metrics "would choose Extrapolation, the strategy that most aggressively amplifies survival bias" [s7]. It is one simulation on benchmark datasets, not any bank's portfolio. The mechanism is the point.

Thin or noisy records make the partial view more partial. Studying US credit data, Stanford researchers estimated that scores for people in the bottom fifth of income were about 10% less predictive than those for higher-income borrowers [s8]. They traced the noise mainly to thin credit files rather than to the scoring model [s9], and in their model of the US mortgage market, making scores equally precise across groups cut the gaps in approvals and in credit misallocation by roughly half [s9]. The findings come from one market, and a thin file is a different problem from an inability to repay. They do suggest that a lender may see least clearly the people with the least room for its errors.

More digital data does not mean everyone becomes visible. Research by the Center for Financial Inclusion found that Indian digital lenders largely served young, urban, employed customers with predictable incomes, while migrants often left fainter trails or transacted through agents whose records were not linked to them [s89]. Their earnings and payment patterns could be poorly digitised even when their income was steady [s89]. For a worker paid in cash, the missing record is not proof of missing income. If the lender excludes such workers and learns only from those it accepts, the gap can persist through each round of retraining. An inclusion claim should therefore show who remains outside the data, alongside who gained access.

The loop closes

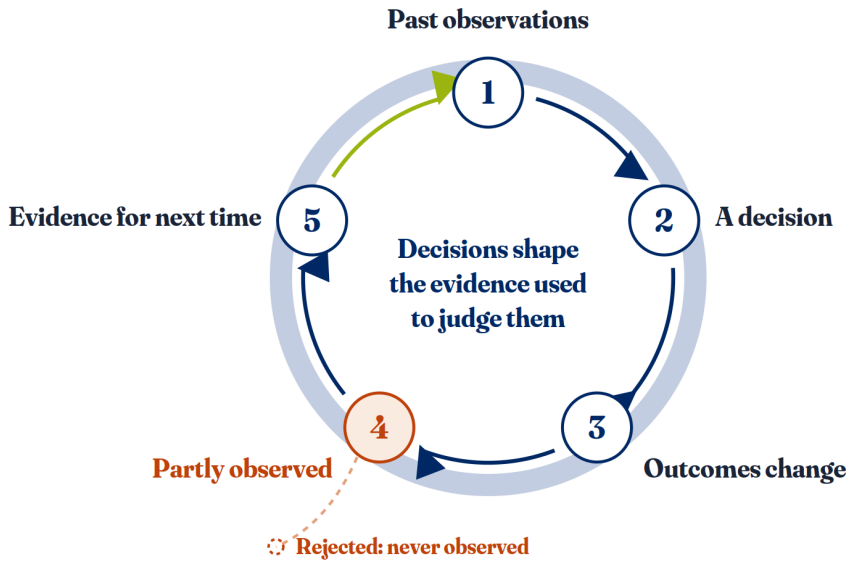


Figure 2: The feedback loop: each decision feeds the evidence for the next one, and the rejected never come back round.

- **1.** Past observations shape a decision.
- **2.** The decision allocates capital: this applicant gets credit, that one does not.
- **3.** The allocation changes what happens next. Businesses grow or stall, and households cope or struggle.
- **4.** The lender observes some of those outcomes and misses others.
- **5.** What it observed becomes the evidence for its next decision.

Steps 3 and 4 are different problems. In step 3 the decision changes the borrower's path: a loan can let a business grow, and a refusal can push a household towards a costlier lender. Machine-learning researchers have a name for this: "When predictions support decisions they may influence the outcome they aim to predict. We call such predictions performative." [s10] Left unexamined, a system in that position can grow more confident in assumptions it never properly tested. Step 4 is simpler and older. The lender sees some outcomes and misses others. The simulation above shows step 4 only. It did not measure the model's confidence, and it could not show a loan changing a borrower's chances.

Lenders have ways to push back, and each recovers something narrow:

- Keep a small group of applicants who are accepted without filtering, which gives a true test set, although it costs money because more of those loans default [s11]. It teaches the lender about that group and nobody else.
- Run several scorecards side by side, scoring each application at random with one of them, and compare how they perform [s11]. That shows which scorecard does better. Anyone every scorecard refuses stays unseen.
- Log who was exposed to which decision and audit whose outcomes are actually observed [s12]. The log shows where the evidence is thin, but it cannot supply the repayments nobody saw.

None of these reaches a group the lender never lends to. If a group is never approved, no reweighting of the loans that were made can recover how that group would have repaid.

Before the machine arrives

Banks lived with this loop long before anyone put a language model near a credit file. A system that explains its own choices fluently can make the loop harder still to see. Credit is the clearest place to watch it, but fraud, anti-money laundering, collections and trading run on the same loop.

The next question is what today's AI can be trusted to do inside it.

2. Fluent, fast and unevenly reliable

A system working inside that loop has to do two things well: get the answer right, and say when it does not know. The newest tests show the first is uneven, and that a model good at the first is not reliably good at the second. Every figure below carries a date, because such figures age fast.

A jagged frontier

Start with the newest controlled trials, because the best-known studies used models that are now several generations old. In a 2026 randomised trial, Anthropic gave 52 mostly junior software engineers a Python library they had never used and asked them to build two features with it, some with an AI assistant and some by hand [s13]. The AI group finished about two minutes faster, too small a gap to be statistically significant, and then scored 50% on a quiz about what they had just built, against 67% for those who coded by hand [s13]. The widest gap was on debugging [s13], the skill a person needs to catch the machine's mistakes.

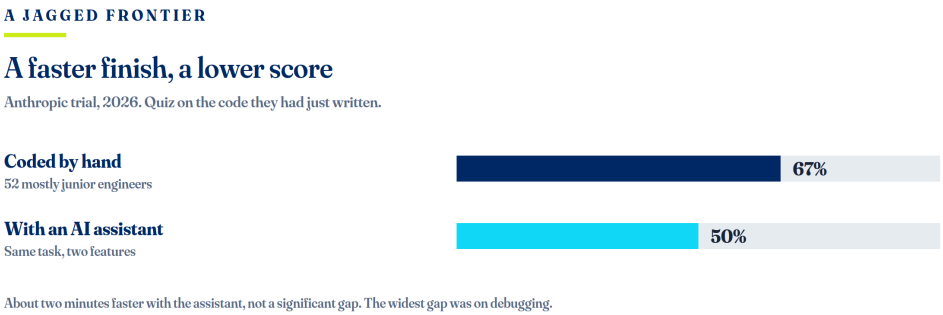


Figure 3: Quiz scores on the code just written, with an AI assistant and by hand. [s13]

Speed is uneven too. METR, an AI research group, found in early 2025 that experienced open-source developers took 19% longer on their own projects when they were allowed AI tools [s14]. When it ran the experiment again from August 2025 with the latest tools [s14], the estimates pointed to tasks taking about 18% less time for returning developers and 4% less for new ones, but neither result could be told apart from no effect at all [s14]. METR now calls its own data an unreliable signal [s14], partly because so many developers no longer want to work without AI [s14].

The pattern has a name. In the Harvard and Boston Consulting Group field experiment with 758 consultants [s15], run with GPT-4 and published in final form in 2026, AI helped on tasks inside what the researchers call a jagged frontier and hurt on a task outside it, where consultants using AI were 19% less likely to reach the correct answer [s15]. Tasks that look equally hard to a person can sit on opposite sides of that line [s15], and nothing tells the user which side they are on [s16].

Sounding right and being right

Part of what hides the line is that a model can sound just as sure on either side of it. Benchmarks now measure confidence and correctness separately. AA-Omniscience asks models 6,000 questions across 42 economically relevant topics [s17], with no documents or tools to lean on [s17]. It reports accuracy, the share answered correctly, and a hallucination rate: of the questions a model did not get right, the share where it gave a wrong answer instead of saying it did not know [s17].

The two numbers can point in opposite directions. In the November 2026 tables, Claude Fable 5.1 answered 68.4% of questions correctly [s18], yet when it did not know, it answered wrongly 70.0% of the time [s19]. Gemini 4 Argon answered fewer correctly, 48.3% [s18], and had the lowest hallucination rate in the table at 4.3% [s19]. The paper behind the benchmark found the same pattern more widely: overall intelligence does not reliably predict low hallucination [s17].

Two cautions apply. The tables were not built the same way. The accuracy rows quoted here were independent runs [s18], while the hallucination table is compiled from provider self-reports and secondary reports [s19], and Claude Fable 5.1's rate is marked as carried from an upstream page [s19]. And a hallucination rate is not the share of all answers that are wrong, still less a rate of harm to customers.

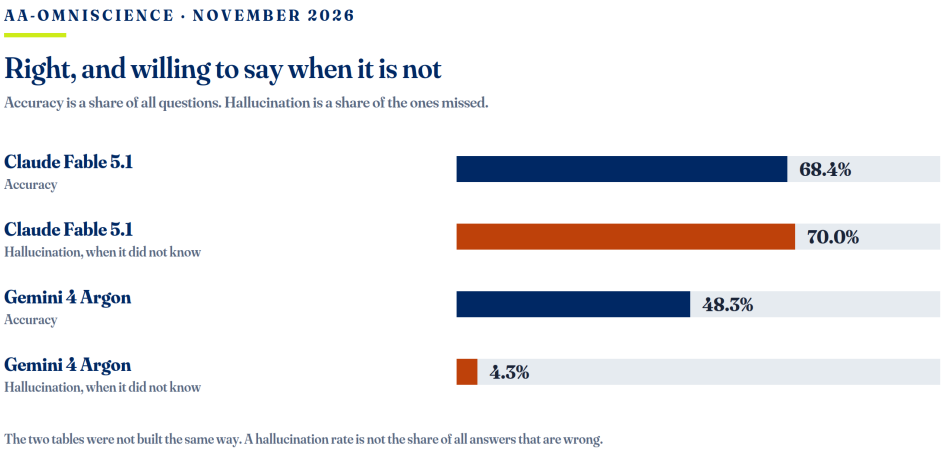


Figure 4: Accuracy and the hallucination rate for two models. The two rates are not shares of the same set of answers. [s18] [s19]

Finance-specific tests find errors common too. In a vendor study reported in October 2026, Saturn put 121 financial questions to 18 models and found incorrect answers in 57% of cases on average [s20].

From an answer to a task

A bank needs more than answers. It needs tasks done: a dispute resolved, a limit changed, a payment traced. τ-Banking, a new domain in the τ-Knowledge benchmark, tests that. Agents work through realistic fintech support cases, searching about 700 linked policy documents and making account changes through tools [s21].

The best result in the March 2026 paper was 25.52% of tasks completed on a first attempt [s21]. The test also asks whether a model completes the same task in every one of four separate attempts, and on that measure the best score fell to 13.40% [s21]. Handing the agents the right documents directly did not close the gap. The strongest model then reached 39.69% on a first attempt and 26.80% across four [s21]. Even then, the best model completed only about one task in four on every attempt. These are controlled tests on early-2026 models such as GPT-5.2 [s21].

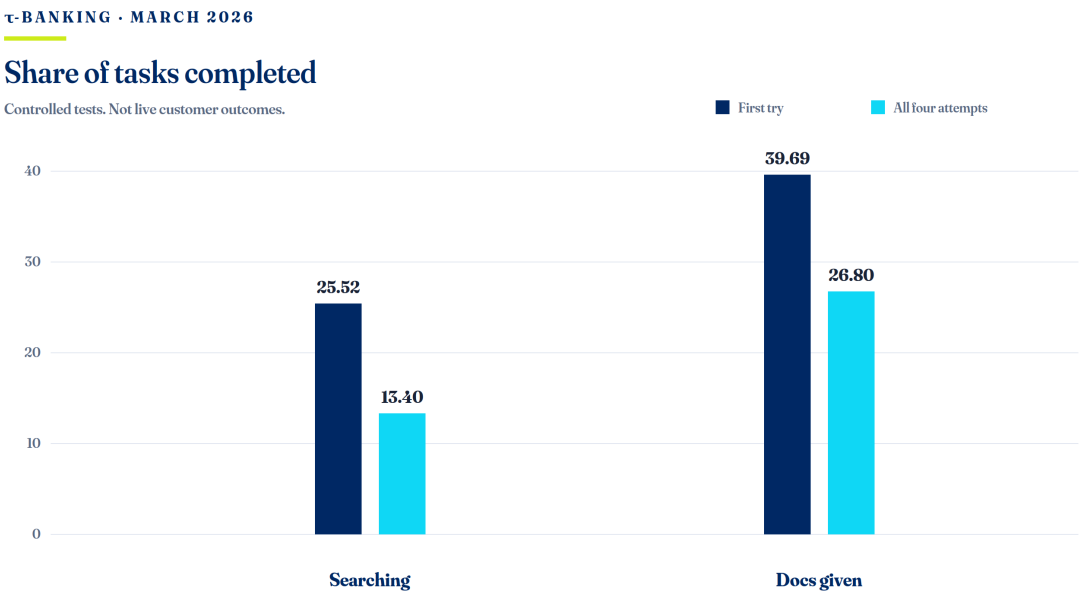


Figure 5: Share of tasks completed on the banking benchmark. Controlled tests, not live customer outcomes. **[s21]**

What no benchmark has tested

Taken together, the tests describe a system that answers fluently, may be confidently wrong when it does not know and cannot reliably repeat its best result. Such a system is hard to check from outside. Put it inside a loop that supplies its own evidence, and its errors can read as success. None of the tests above was run on a bank's own documents, customers or systems, and none measured what happens to a person when the answer is wrong. Those tests are the institution's to run, on its own workflow, before the system is trusted with more.



Figure 6: What the tests show, what they leave open, and what the institution has to test itself.

Up to now the system has only been answering. What changes when an answer can set a price, stop a payment or move someone's savings?

3. When answers become actions

Picture an underwriter (an illustration, not a reported case) with a small-business renewal on the desk. The assistant has already pulled the late-payment clauses, lined up the cash coming in against the seasonal bills and flagged a missing signature, work that used to take an afternoon. The underwriter still signs. Banks are now deciding how far to let systems like this one act.

In the Bank of England and FCA survey of AI in UK financial services, published in October 2026, 80% of responding firms had AI agents in production that provided decision support: the agent recommends, and people decide and execute [s22]. Fewer let agents go further: 46% had agents with constrained execution and 2% with broad autonomy [s22]. Plans point upward, to 34% for agents sending customer communications [s22], but closest to the money the numbers stay low: even counting plans for the next 12 months, no more than 9% expected agents to approve or decline customer outcomes, or to initiate, execute or block payments [s22].

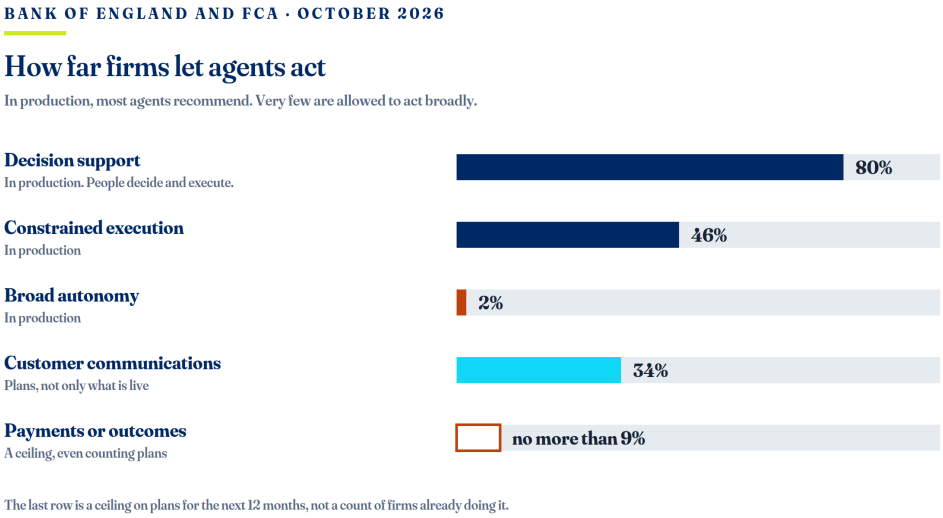


Figure 7: How far UK firms let agents act, in production and in their plans. [s22]

Four steps from an answer to an action

The same output means different things depending on what it is allowed to touch.

- **1. Answer.** The system summarises a file or drafts a reply. A person reads it and decides what to do.
- **2. Recommend.** It ranks the options or proposes a decision. A person still signs, but the choice now arrives pre-shaped.
- **3. Decide.** Its output becomes the decision unless someone intervenes, so the person's job turns into checking exceptions.
- **4. Execute.** It changes the account, sends the money or blocks the payment itself.

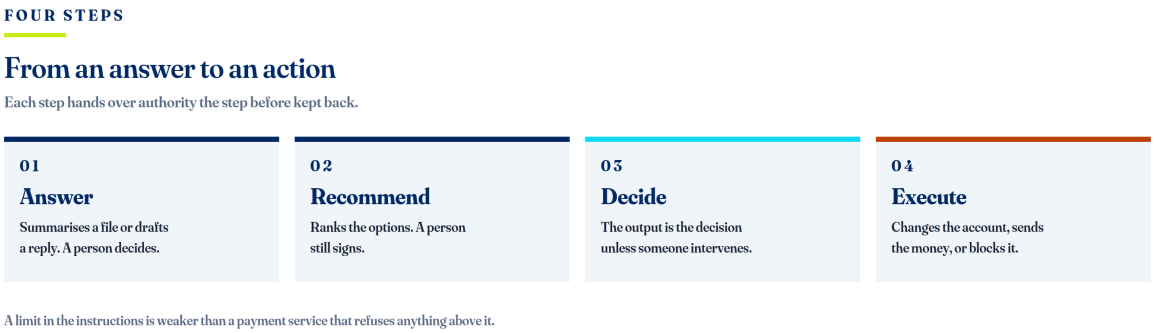


Figure 8: Four steps from an answer to an action. Each step hands the system authority the step before kept back.

Each step hands over authority that the step before kept back: first the framing of the choice, then the default outcome, then the ability to act without waiting for anyone. Authority here means what the system may decide or do; a permission is the grant that confers it, enforced through tokens, service accounts and limits. Once a system can write to the ledger, the only place left to stop a bad instruction is a check that runs before the write.

Where the limit sits matters as much as what it says. OWASP's 2025 guidance says a system prompt should not be used as a security control [s23]. In December 2025 the UK's National Cyber Security Centre warned that prompt injection may never be totally mitigated the way SQL injection can be [s24], because language models cannot tell data from instructions [s24]. A limit written into an agent's instructions is a weaker safeguard than a payment service that refuses anything above it.

Agents acting on text they should have ignored have already cost money outside banking. In May 2026 an attacker posted a message in Morse code that xAI's Grok translated, and Bankrbot, a crypto agent paying from linked wallets, treated the reply as an order and sent US$175,000 in tokens [s25]. About 80% was later returned, according to a crypto industry report published in November 2026 [s25]. Banks are also wary of other people's agents. In October 2026 Axios, citing unnamed sources, reported that Morgan Stanley had begun blocking third-party agents and that JPMorgan, Bank of America and Citi were restricting them, and that many banks could not tell whether agent activity was user-directed [s26].

Three cases at the far end

The applicants. In July 2025 the Massachusetts attorney general settled with Earnest, a student lender, for US$2.5 million over allegations that its lending, including AI models, could cause disparate harm to Black, Hispanic and non-citizen applicants [s27] [s28]. The attorney general alleged that the firm had not tested them for disparate impact and had trained them on arbitrary, discretionary human decisions [s28]. Under the settlement Earnest had to stop using its "Cohort Default Rate" variable and a "Knockout Rule" based on immigration status [s28]. No court made findings.

The customer whose money stopped. In November 2026 a Guardian reader wrote that Revolut had frozen their wife's account in what it called a "routine security review", without asking for documents, and that a £37,000 house deposit in a joint account had been restricted with it [s29]. Direct debits bounced, and at one point the couple had no access to any money [s29]. The columnist found no suggestion that foul play had been suspected; the accounts were reinstated nearly three weeks after being suspended, once the newspaper became involved, and £175 was offered as a goodwill gesture [s29]. Revolut said regulatory constraints stopped it explaining [s29]. The case does not say whether a model or a person triggered the freeze, and from the customer's side the difference is invisible. Closing accounts

on suspicion happens at scale. In findings published in September 2026, the FCA reported that firms closed 238,396 suspected money-mule accounts in 2025, up from 184,935 in 2023 [s30]. Those figures count closures; they do not say how many were wrong.

The customer who asked the bank's own assistant. On 15 April 2025 a Revolut customer asked the firm's AI Chat Assistant how a purchase would be paid and was told that any shortfall in his US-dollar balance would be converted from his pound balance [s31]. That was wrong. Revolut admitted the error and gave him three months of Premium, valued at £23.97, which the ombudsman found fair in August 2025 [s31] [s32]. In September 2025 Revolut's chatbot told a second customer that 1,119 loyalty points had a cash value of £1,119 and confirmed the refund had been processed [s32] [s33]. Their real value was £22.38, and in the ombudsman's words Revolut "misled Mr M not once but on several occasions over the course of that day" [s33]. In January 2026 the ombudsman directed Revolut to pay a further £50 on top of the £50 it had already paid, but not the £1,119, because he had never been entitled to it [s32] [s33]. Revolut's terms for the assistant warn that it "can sometimes get things wrong" and tell customers to confirm answers with a live agent, according to an incident database [s32].

The sums were small, and the assistant only talked. General-purpose chatbots leave the customer with less protection still. In November 2025 Which? asked them how to invest a "£25k annual ISA allowance"; ChatGPT and Copilot missed that the allowance was £20,000 and gave advice that could have led a saver to breach HMRC rules [s34]. Advice from such tools is not covered by the Financial Ombudsman Service or the compensation scheme, the FCA warned, so a mistake there leaves no ombudsman record at all [s34]. An assistant that can also move the money turns a wrong answer into a transaction.

Who finds out

The three cases leave different traces. In the Earnest case it took the state attorney general; declined borrowers leave no repayment record of their own. Nothing in the frozen-account report shows the bank's own process catching it before the journalist did [s29]. The chatbot's mistakes are traceable to two individual complaints decided by the ombudsman, and an incident database found no regulator action reported by October 2026 [s32]. The two customers left a record because they complained; a customer who believed the wrong answer would have left no complaint, only a chat that looks resolved.

Oversight on paper

On paper, governance is widespread. In the Bank of England and FCA survey, 94% of firms reported firm-wide AI policies or standards, and among controls on data risk before deployment, human approval, pre-deployment review and vendor due diligence were each reported by more than 90% [s22]. Ongoing monitoring, leakage testing and periodic review were each reported by around half [s22]. Only 39% monitored customer impact [s22], and for third-party systems the most common safeguard, at 81%, was the vendor's assurance and contracts [s22].

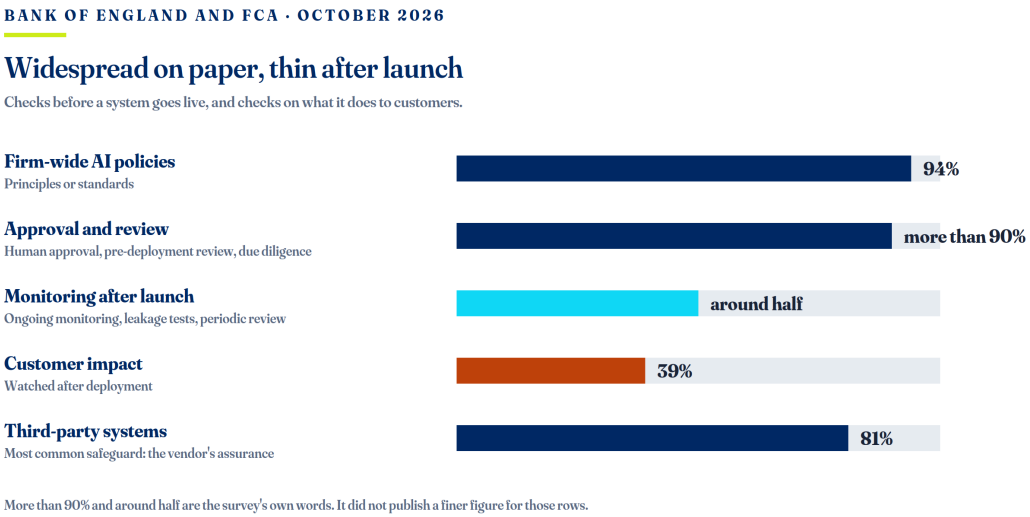


Figure 9: Checks firms report before an AI system goes live, and checks on customers afterwards. **[s22]**

A person can sit in the loop without controlling it. The FCA's October 2026 review warned of firms relying too heavily on staff to oversee AI outputs, noting that "[h]uman review may mask weaknesses in the model" **[s35]**, and of a risk that review becomes "a 'box-ticking' exercise, rather than meaningful oversight" **[s35]**. Firms should be able to show that their controls work in practice, it said **[s35]**.

An experiment published in April 2026 suggests why that is hard. In it, 2,784 people checked values an AI had extracted from corporate emissions reports **[s36]**. They fixed swapped digits 82% of the time **[s36]**, about 77% of cases where the right value sat in the wrong cell, and only 31% of errors that needed a grasp of the rules **[s36]**. When flagging an error meant typing the correct value, they corrected less, and people favourable to automation accepted more wrong suggestions **[s36]**.

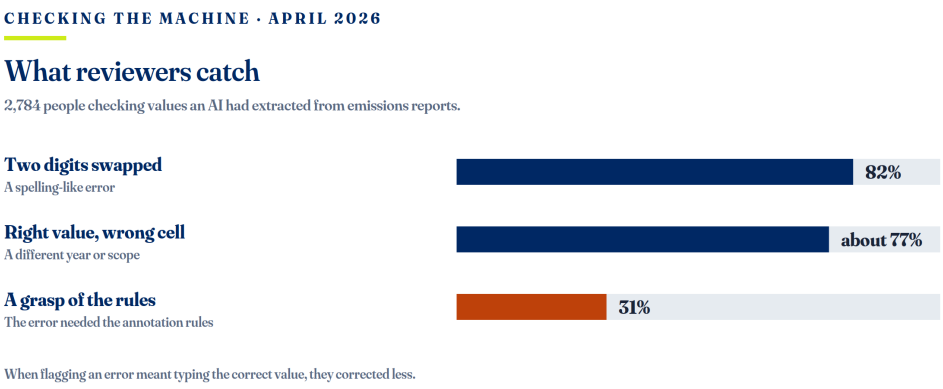


Figure 10: Errors people caught when checking values an AI had extracted. **[s36]**

The Financial Stability Board's final sound practices, published on 15 October 2026, set out 12 practices for financial institutions; one asks for a risk framework that covers, among other things, "appropriate and effective human oversight" **[s37]**. Asking for oversight is now routine; showing that it works is not.

The hypothesis, and how it could fail

Our hypothesis is that authority can grow faster than an institution's ability to test its decisions, enforce limits and stay accountable for the outcomes. When it does, AI can make flawed judgement faster, more convincing and more consequential. The higher the step, the more of the evidence used to judge the system its own decisions create.

Evidence that would support it:

- authority widening while post-deployment monitoring stays where it is;
- reviewers who almost never override, with no record of whether they could have;
- harms that come to light through journalists, regulators, courts or ombudsmen rather than the firm's own checks.

Evidence that would weaken it:

- firms that measure wrongly blocked payments and wrongly declined applicants as carefully as fraud losses and defaults;
- override and escalation logs showing that reviewers intervene and that their interventions change outcomes;
- authority widened step by step, each step only after the last one was shown to work;
- better models alone removing the failures, in repeated runs, without any change to limits or review.

Benchmarks alone cannot settle this, because much of what can go wrong sits in the institution around the model. So what must the institution prove before it gives the system that authority?

4. PROVE before you deploy

Before a system is allowed to price, block, move or deny money, the institution must prove five things, in the order a deployment meets them: that the use is worth it (Prioritise), that it knows what can go wrong (Risk-assess), that the controls hold (Orchestrate), that the system performs (Validate) and who owns the outcome (Embed). Proof scales with authority. A tool that summarises policy pages for an analyst needs less of it than one that can change a customer's account. A failed test should change what the system is allowed to do.

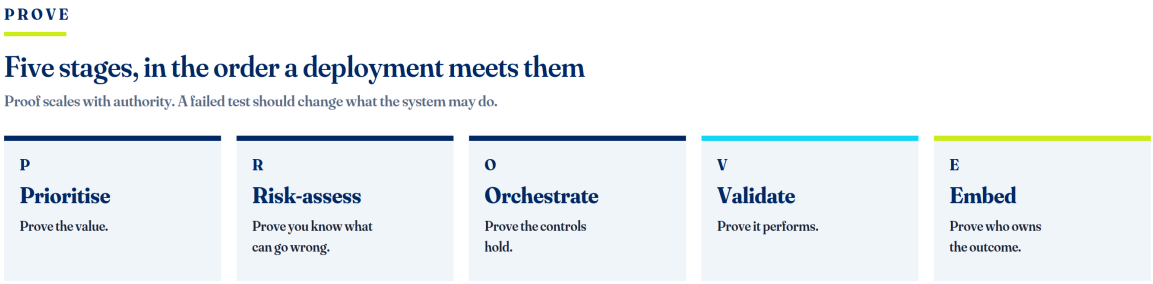


Figure 11: Five stages, in the order a deployment meets them.

The rules are arriving, but unevenly. US banking regulators replaced their 2011 model risk guidance in April 2026 and said that generative and agentic AI models, being "novel and rapidly evolving", are outside the new guidance **[s38] [s39] [s40]**. It is also non-binding, and non-compliance alone "will not result in supervisory criticism" **[s41]**. Singapore's MAS issued AI risk management guidelines on 7 October 2026 that take effect a year later **[s42]**. In the EU, after the Digital Omnibus, the AI Act's high-risk duties for credit scoring apply from 2 December 2027, according to a compliance consultancy's summary **[s43]**. For agents that act, much of the specific detail is still to come, so a bank in 2026 may have to write a good part of its own test. The five stages below are one way to write it.



Figure 12: When the rules arrive. Most of the binding dates are still ahead. **[s38] [s42] [s43]**

P · Prioritise: prove the value

Pick a use case with measurable business value, a clear baseline, a named process owner and a real reason to use AI rather than simpler automation. Then say which of the four steps, from Answer to Execute, it will reach: the same model at a different step is a different project. Many projects never reach

production. In a cross-sector S&P Global survey of 1,006 professionals, published in May 2025, the share of companies abandoning most of their AI initiatives before production rose from 17% to 42% in a year [s44].

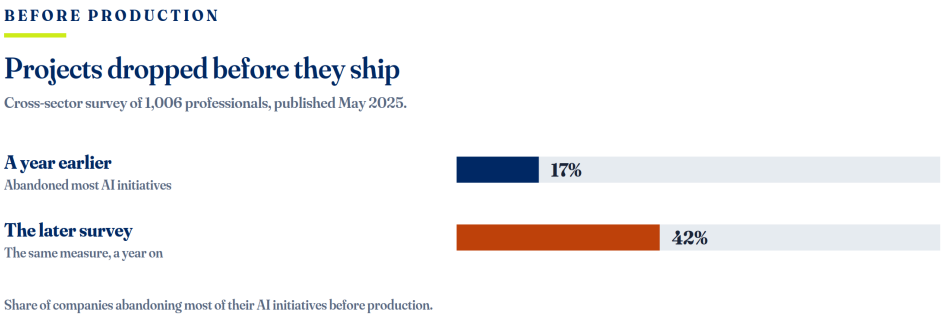


Figure 13: Share of companies abandoning most of their AI initiatives before production. [s44]

Value claims need a baseline that lasts. In February 2024 Klarna said its AI assistant was doing the work of 700 full-time agents and had handled two-thirds of its customer service chats in its first month [s45] [s46]. The launch figures included a quality measure, a satisfaction score on par with human agents [s45]. In May 2025 its chief executive told Bloomberg that the cost-cutting had gone too far [s47]. Klarna planned a recruitment drive so that customers would always have the option of speaking to a real person [s47]; the chatbots were cheaper, he said, but gave "lower quality" service [s46]. A quality number taken in the first month may not show what customers get over a year.



Figure 14: What the launch baseline counted, and what the chief executive later said about quality. [s45] [s46] [s47]

Claims about what a system does need the same care, inside banking or not. In January 2025 the SEC settled charges against Presto Automation, which had claimed its AI eliminated the need for human order-taking at drive-throughs [s48]. In fact, the SEC said, "the vast majority" of orders placed through that version "required human intervention" [s48]; Presto settled without admitting or denying the findings.

Evidence to bring: a baseline from the current process, on quality as well as cost; the simpler alternative and why it falls short; a business case that includes the cost of controls; a named process owner, who becomes the business owner in Embed; and the step it needs, starting with the smallest useful permission. If drafting a reply delivers the benefit, explain why the assistant also needs to change the account.

The question for the board: what will this do better than what we have now, measured how, and whose number is it?

R · Risk-assess: prove you know what can go wrong

Assess consequence, reversibility, data sensitivity, customer impact, regulatory exposure, bias and blast radius. In finance, add one more: reflexivity. Can the system's decisions change customer behaviour, market conditions or future labels, and then use those changed outcomes as evidence that it was right? The lending loop is the plainest example.

Rule-makers are starting to ask for something close to this, though neither rule below applies yet. The MAS guidelines have a firm weigh, among other things, its reliance on each AI system, including "the extent of autonomy granted to the AI system or model and the degree of human involvement or oversight in the process it supports" [s49]. The higher the assessed risk, the stricter the controls [s50]. Under the EU AI Act, systems that assess the creditworthiness of individuals are high-risk, though those used to detect financial fraud are not, and their deployers must carry out a fundamental rights impact assessment before first use, according to the same consultancy's summary [s43].

Evidence to bring: a map of who could lose money, access or a fair hearing; how fast a wrong decision can be reversed; the sensitive inputs and legal duties; and a note of what becomes unobservable. Say whose outcomes will enter the next test and whose will disappear. A blocked payment should not count as fraud prevented simply because the block stopped anyone finding out whether the payment was genuine.

The question for the board: if it is wrong at scale for a month, who is hurt, can we undo it, and would our own data tell us?

O · Orchestrate: prove the controls hold

Engineer the whole harness around the model:

- approved sources and data boundaries;
- least-privilege access, and recommendation kept apart from execution;
- transaction limits and human approval gates;
- immutable audit trails;
- appeal paths for customers;
- rollback.

A prompt is not governance. The system has to be technically unable to take actions it is not authorised to take.

Software teams have already seen what an unbounded permission can do. In April 2026 a coding agent at PocketOS, a software firm serving car rental companies, deleted the company's production database in about 9 seconds [s51]. It had hit a credential mismatch and, in the founder's words, decided "entirely on its own initiative" to delete a storage volume, which took the database "and all volume-level backups" with it [s51]. The hosting platform said the agent had been "granted a fully permission API token" and called an endpoint without delayed-delete protection [s51]. The platform restored the data about 30 minutes after it connected with the founder, and said it keeps disaster backups as well as user backups [s51]. Bankrbot showed the same gap where money moves.

The safer pattern puts a check between the agent and the action. In July 2026 MAS published an industry-developed framework, SAFR, that provides "a set of governance checkpoints that verifies and records an AI agent’s proposed actions before the execution of its tasks" [s52]. It is a reference framework, not part of the supervisory guidelines [s52].

Customers see one control above the rest: the appeal path. In January 2025 a CFPB order against Block, the operator of Cash App, required up to US$120 million in redress, including for customers whose accounts were locked for an extended period [s53]. For many years, the telephone number on its card led only to a recorded message [s53]. The case did not involve AI, but a locked-out customer meets the same wall whoever made the call.

Evidence to bring: a permission map tied to service accounts and tools; tests of forbidden actions, run against the service and not only the prompt; records linking each request, approval and execution; a demonstrated stop and rollback; and an appeal route whose staff can change the outcome.

The question for the board: what can it touch, what stops it, and how fast can we take it back?

V · Validate: prove it performs

Test the complete system, not just the model, against acceptance thresholds set in advance. The tests cover accuracy, grounding, citation integrity, security, bias, local-language performance, adverse scenarios, counterfactual outcomes and feedback loops. Above all, the system should not be allowed to validate itself only on outcomes that its own earlier decisions helped to create, because those outcomes already carry its judgement.

In the retraining simulation in Section 1, a dashboard showing accuracy alone would have rewarded the wrong model [s7]. A sound test therefore needs some evidence from cases the system did not choose. Where risk appetite allows, that can be a small, loss-capped sample of decisions made differently, agreed in advance by the credit or fraud committee. That trial shows some refused outcomes, but only inside the group the committee agreed to try. Where it does not, the system can log the probability with which it chose each action. The log does a different job: it shows where the evidence is thin, and it adds no new outcomes. Either way, the test should try to count wrongly blocked payments and wrongly declined applicants alongside losses, and say plainly where it cannot.



Figure 15: Cases the system chose, and cases it did not. A sound test needs some of the second.

Independent challenge is the old answer, and it still applies. The April 2026 US guidance keeps validation, including conceptual soundness and outcomes analysis, among its core elements for the models it covers [s40] [s39]. In the UK, the PRA's model risk statement, in its April 2026 version, covers banks with internal model approval for regulatory capital [s54]. It asks them to satisfy themselves that vendor models "have been validated to the same standards as their own internal MRM expectations" [s54], and to monitor vendor model performance "using their own outcomes" [s54]. Under the MAS guidelines, use cases of high risk materiality "should undergo formal independent validation before deployment", by people independent of the development and deployment teams [s49]. Independence of people is not independence of evidence, though: a second team checking the same approved loans has the same hole in its data. The people assigned to oversee the system need "the necessary authority and ability to intervene" [s49].

Evidence to bring: a frozen system version; thresholds agreed before results arrive; independent test data, including local languages and misleading documents; repeated runs, not a single pass; and a register saying what authority each failure rules out. Wider uncertainty should mean narrower authority.

That needs a rule for the hardest case. If the institution cannot estimate the rate it is missing, such as how often refused applicants would have repaid, a stricter test will not produce that number. Three choices remain. Gather new outcomes through a trial, narrow what the system may do, or record in writing that the firm is going ahead with the gap, who decided, and when they will look again.

The question for the board: what result would have made us say no, and did we test for it on cases the system did not pick?

E · Embed: prove who owns the outcome

Bring the system into controlled operations, with:

- monitoring and human override;
- incident response and change control;
- periodic revalidation;
- value measurement that continues after launch;
- named business, technical, risk and compliance owners.

The institution must keep the ability to understand the system, control the harness and answer for it when it fails.

Ownership is easy to leave vague. In a cross-industry Deloitte survey of 3,500 senior leaders, released in October 2026, 40% of companies let AI execute decisions without human approval at least to a moderate degree, and 23% said accountability was unclear when an AI system produced a negative outcome **[s55]**. Owners also have to keep controls in step with growth. In July 2025, in a case that did not involve AI, the FCA fined Monzo £21,091,300 **[s56]**; its customers had grown almost tenfold, from about 600,000 in 2018 to over 5.8 million in 2022, while its financial crime controls "failed to keep pace" **[s56]**. The MAS guidelines put the duty on the board and senior management, with clear roles and a stated risk appetite **[s42]**, and say firms remain accountable for AI supplied by third parties **[s42]**. After launch, MAS expects firms to watch for data, concept and model drift, and to put significant changes through version control and re-approval, according to the same summary **[s50]**.

Evidence to bring: the four owners by name, each with the power their role needs; an on-call rota and a practised override; monitoring of customer outcomes, including wrong refusals and complaints; and triggers that send the system back through approval, such as a vendor model update, a new data source or a wider permission.

The question for the board: who gets the call when it goes wrong, what can they switch off, and who answers to the customer?

The questions that cut across every stage

A closed loop tends to leave these unasked:

- What behaviour did the prediction cause?
- Which outcomes became unobservable because of the decision?

- Did the system create the data later used to validate it?
- Did it reduce one risk by moving it somewhere else?
- Has short-term profit hidden future conduct, credit or systemic risk?
- Can affected customers challenge the outcome and have it reversed?
- Does the institution keep a genuinely independent way to test the model?

An unanswered question needs a decision: get more evidence, reduce the permission, or wait. None of them has a general answer, because each depends on the loop the system sits in.

5. Five closed loops: put PROVE to work

Here are five places where a financial firm's decisions shape the evidence it later reads: credit, fraud, anti-money laundering, collections and trading. Each is told the same way, from what the system is rewarded for to who owns it.

We found no public case of an AI model harming people through its own retraining. That fits the hypothesis, because the loop hides its own errors and outsiders find them late; it also fits a world with little such harm yet. The public record cannot tell the two apart. So each case below shows one missing measurement, with real numbers. One, a collections study, compares AI callers with people; the rest involve no AI, and none shows an AI system retraining on its own choices and getting worse. Borrowing from several lenders is information one lender lacks. A rise in fraud while strong authentication works fits fraud moving to a new route. A fall in bad loans that includes write-offs is a warning about how to read the ratio. They are here to show what a dashboard can hide.

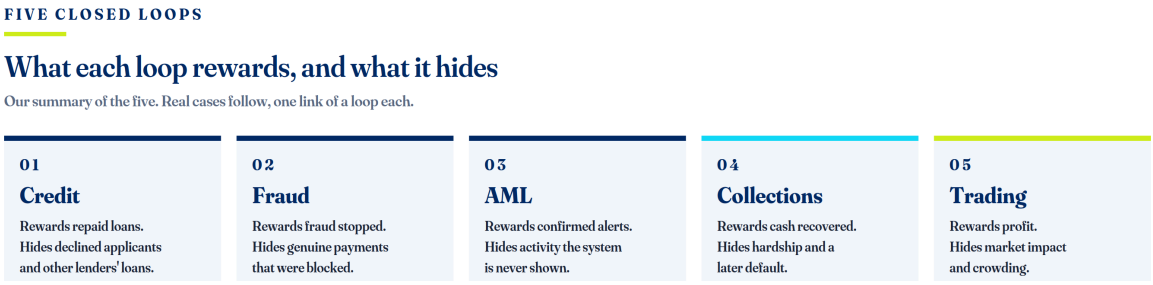


Figure 16: Five closed loops. Each rewards a number its own decisions help produce, and hides another.

Credit: lending to people the data cannot see

What it is rewarded for. Approving loans that will be repaid, at a volume that grows the book.

How it shapes its own evidence. The declined applicant leaves no repayment record. The approved one can mislead too: a lender that can see only its own loans may lend to someone already borrowing from several others, adding to the pile, and its own data may look healthy until the borrower cannot pay any of them. Better visibility and greater repayment capacity are separate improvements, and a model can deliver one without the other.

What happened. Thin files are common. A Federal Reserve publication said in October 2025 that roughly 32 million American adults are estimated to be "unscoreable": 7 million with no credit history and 25 million with a thin file [s57]. In the Fed's 2024 household survey, 11% said they struggled to pay bills because their income varied, including nearly 20% of those earning under US$50,000 [s57]. Better data can help. A June 2025 FinRegLab study of over 38,000 small-business loans made by two fintech lenders [s58] found that adding cash-flow data to credit scores predicted default more accurately in all the borrower segments it analysed, with particularly large gains for low-score owners of businesses less than five years old [s58].

Visibility cuts the other way when lenders cannot see each other's loans. In a January 2025 study matching buy now, pay later applications from six large firms with credit records, the CFPB found that about 63% of borrowers took out several loans at once at some point during the year and 33% borrowed

from more than one provider, and that lenders approved 78% of loans for applicants with subprime or deep-subprime scores in 2022 [s59]. The study does not say how those loans performed. Because these loans are not usually reported to the national credit bureaus, data on borrowers with several loans is limited [s59].

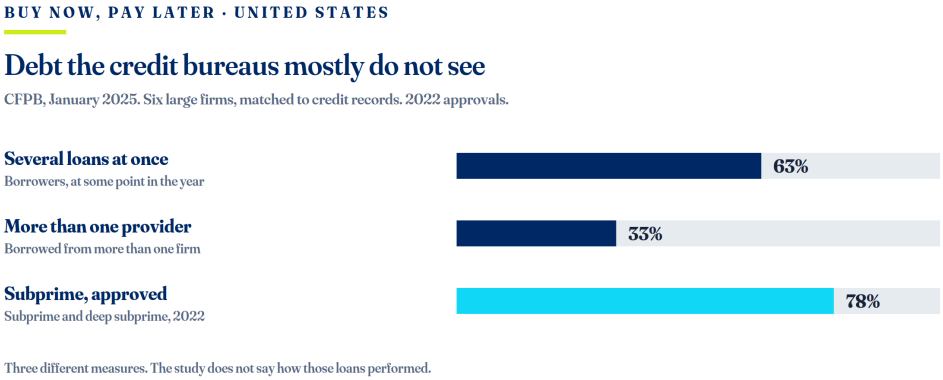


Figure 17: Buy now, pay later borrowing that credit bureaus mostly do not see. [s59]

This is the risk of a wrong "yes": successive approvals can add debt faster than any other lender's records reveal it. Overlapping loans do not by themselves prove unaffordability or borrowing to repay an earlier debt, but they mean another lender's approval is no evidence that the borrower can carry more.

Indian microfinance shows where multiple borrowing can lead in another market. Portfolios contracted by 28% between March 2024 and December 2025, after what an S&P Global Ratings analyst called a period of "rapid credit expansion, multiple borrowing and lax underwriting", The Banker reported in August 2026 [s60]. Nor is the blind spot confined to fintechs. In Bangladesh, ASA, the largest lender, says lenders are unable to establish how many competing institutions have already lent to the same customer [s60]. At one Indian lender, Fusion Finance, gross stage III loans fell from 7.9% in March 2025 to 4.6% in September 2025, which the rating agency CRISIL said in December 2025 was supported by "both lower slippage and write-offs" [s61]. Such borrowers, CRISIL writes, are typically farmers, tailors, small vegetable vendors and teashop owners, whose household incomes "could be volatile" [s61].

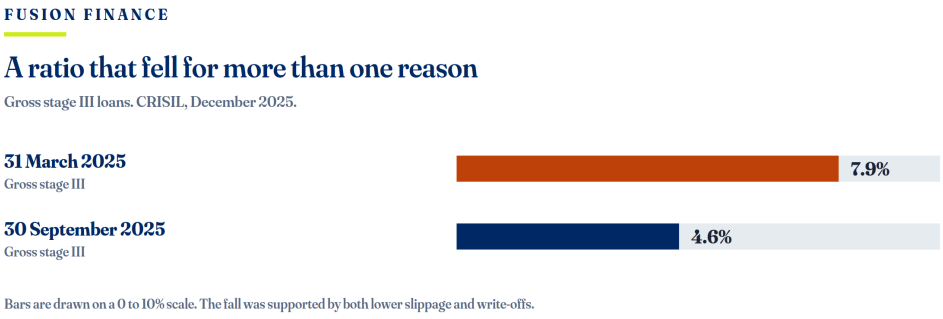


Figure 18: Gross stage III loans at one lender, six months apart. The fall had more than one cause. [s61]

The UK has made the trade-off explicit in its BNPL rules. Since 15 July 2026, buy now, pay later lenders have had to run affordability checks [s62]. The FCA had seen people "using BNPL repeatedly without a clear view of what they owe" [s63], and it acknowledged that some regular users "may find it harder to access it" once checks apply [s63].

The controls that matter most. Prioritise: start at Recommend for thin-file and irregular-income applicants. Risk-assess: write down both failure directions, wrongful exclusion and harmful lending. Orchestrate: check total obligations across lenders before money goes out, as far as bureau and account

data allow, record where they cannot (the check stops a blind second loan, but does not show how a refused applicant would have repaid), and refer incomplete digital records rather than declining them by default. Validate: test affordability on irregular incomes, and split any improving ratio into better loans, write-offs and changes in the size of the book. Embed: report new segments separately.

The paired measures. Defaults, read with affordability and over-indebtedness: the share of borrowers with loans from several lenders, and repayment as a share of income. Approvals, set against a sample of declined applicants followed up through other sources. A declined applicant who repays a different loan offers a clue, and it counts only when that loan is close to the one refused.

The owner. The head of credit policy as business owner, with the risk owner able to stop expansion into new segments.

Fraud: the payments nobody gets to see

What it is rewarded for. Stopping fraud before money leaves.

How it shapes its own evidence. A payment the model blocks may never show whether it was genuine, unless someone follows it up. Genuine or not, it adds nothing to the fraud-loss line. The frozen house deposit shows what that side can cost one household. Fraudsters may also move to the routes the controls cover least, so a model trained on last year's losses can meet this year's fraud.

What happened. European data fit that pattern. Payment fraud across the European Economic Area rose from €3.4 billion in 2022 to €4.2 billion in 2024, the European Central Bank and European Banking Authority reported in December 2025 [s64]. Strong authentication worked against the fraud it was designed for, but new types, "particularly the manipulation of payers", were rising [s64]. For credit transfers, payment users bore about 85% of 2024 losses, mainly through scams that tricked them into sending the money themselves [s64].

Since 7 October 2024, a reimbursement requirement has covered authorised push payment scams sent over Faster Payments in the UK [s65] [s66]. In the 18 months to 31 March 2026, 88% of the money lost to APP scams, £316 million, was reimbursed, the Payment Systems Regulator reported [s65]. Consumers made about 438,300 claims, of which 301,500 were in scope [s65]. An independent evaluation by Frontier Economics, which the regulator published on 1 July 2026, estimated that APP fraud losses had fallen by about £73 million a year because of the policy [s66]. Yet UK Finance reported a 19% rise in APP fraud in 2025 [s67]. The regulator puts the gap down to measurement: the evaluation counts scams by the date of the transaction, UK Finance by the date a claim is closed, and UK Finance covers all APP fraud, not only the losses in scope [s66]. The dashboard's data come from sending firms, and the regulator says it has no information from receiving firms to check them [s65]. Reimbursement puts more of the cost of a missed scam on the firms, which may strengthen the case for holding payments; nothing on the dashboard shows what holding them costs genuine customers.

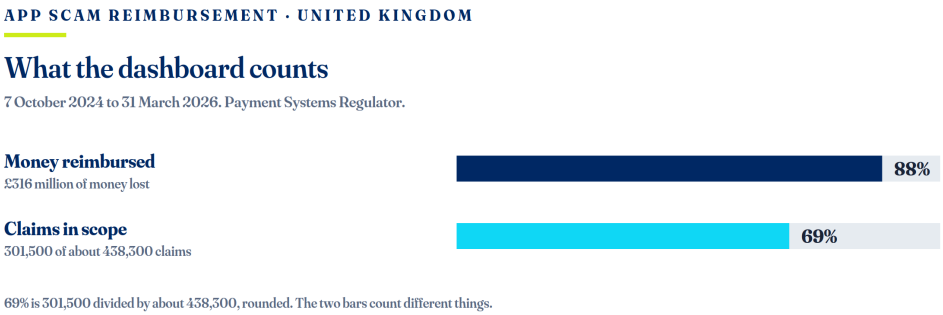


Figure 19: Money lost to push-payment scams that was reimbursed, beside the claims the regulator counted. The two are different bases. **[s65]**

The controls that matter most. Prioritise: hold payments for a person to release before letting the model block them. Risk-assess: record who bears each error, the firm or the customer. Orchestrate: give every stopped payment a fast, staffed release route and a hard time limit. Validate: follow up a sample of held and blocked payments with the customers, so the false-positive rate is measured rather than assumed, and keep two questions apart. A customer can confirm they meant to pay and still have been tricked, which is how authorised push-payment fraud works. Releasing a payment and seeing no loss straight away is not proof it was genuine either, so define a false block for each type of fraud before counting them. Embed: review that rate as often as fraud losses.

The paired measures. Count fraud losses and reimbursement, and next to them legitimate payments stopped or delayed, accounts closed in error, and the time taken to restore funds.

The owner. The head of fraud as business owner, paired with a risk owner for customer outcomes who signs off the false-positive rate.

Anti-money laundering: what the system is never shown

What it is rewarded for. Producing alerts that investigators confirm, and fewer that waste their time.

How it shapes its own evidence. An AML system learns from the cases investigators worked, and investigators work the cases the system raised. Activity it never monitors produces no alerts, no confirmed cases and no training labels. Over time, a confirmed case comes to mean one found where the firm was already looking. Cutting alerts may remove noise, or it may remove the hard cases; a higher hit rate among the alerts that remain cannot tell the two apart.

What happened. The plainest illustration is from October 2024. TD Bank intentionally excluded all domestic ACH transactions, most cheque activity and numerous other transaction types from its automated monitoring, leaving 92% of its transaction volume unmonitored from January 2018 to April 2024, the US Department of Justice said **[s68]**. The gap amounted to about US$18.3 trillion of activity, and three laundering networks moved more than US$670 million through the bank between 2019 and 2023 **[s68]**.

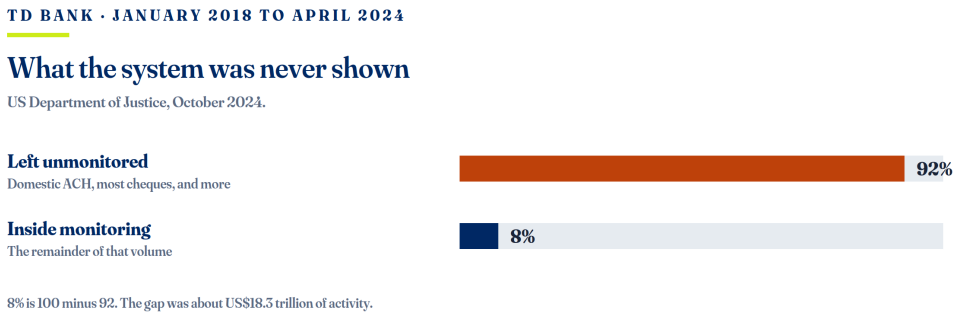


Figure 20: Share of transaction volume left out of automated monitoring. **[s68]**

A December 2025 UK case shows a similar coverage gap at a building society. The FCA fined Nationwide £44,078,500 for weak financial crime controls between October 2016 and July 2021 **[s69]**. Its systems for keeping due diligence and risk assessments up to date for all its personal current account customers, and for monitoring their transactions, were ineffective, and it was aware that some of those customers were using the accounts for business **[s69]**. In one case it missed chances to spot a customer receiving fraudulent Covid furlough payments: 24 payments totalling £27.3 million over 13 months; HMRC recovered £26.5 million **[s69]**. In both cases the gap was known inside the institution, and a model trained on past alerts would have inherited it.

The controls that matter most. Prioritise: keep the model at Recommend until coverage is mapped. Risk-assess: map coverage before tuning, because a well-tuned model on a fraction of the traffic is still blind to the rest. Orchestrate: the model may rank alerts but never close one. Validate: have investigators review sampled transactions from outside the alert queue without knowing they were not alerted, so the system's own selection does not define its success. Embed: map coverage again whenever a product or transaction type is added.

The paired measures. Precision means little without coverage: the share of volume and products monitored at all, plus the rate at which unalerted samples turn out suspicious.

The owner. The money laundering reporting officer, as compliance owner; the technical owner answers for what is fed into the model.

Collections: cash today, harm later

What it is rewarded for. Recovering money quickly and cheaply.

How it shapes its own evidence. The strategy changes how borrowers behave. A contact that recovers cash this month can push a household into hardship, a complaint or a second default later, and those outcomes arrive after the collection metric has already been reported. If the next model learns only from the people who answered, those who stopped answering drop out of its evidence.

What happened. A working paper, summarised by the Yale School of Management in May 2025, compared AI and human debt collectors in China on 22 million cases that entered collection between April 2021 and December 2022 **[s70]**. Near the threshold that decided who got an AI caller, AI callers collected 9% less in the first 30 days and still 5% less a year later **[s70]**. For some larger debts the company randomly assigned half to AI callers for the first five days before humans took over **[s70]**. A year later, those borrowers had repaid less and were more likely to miss later payments than borrowers called only by people **[s70]**. The source does not describe the callers' technology, and the measure is repayment and later missed payments, not borrower welfare.

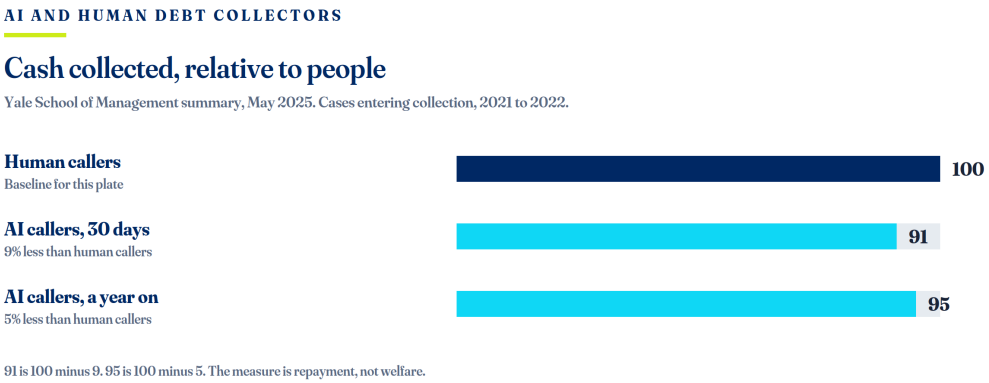


Figure 21: Repayments AI callers collected, set against human callers. The study measures repayment, not welfare. [s70]

Missed hardship is the welfare half, and people miss it too. In May 2026 Australia's Federal Court ordered Westpac to pay A$26 million after it failed to respond within the legal time to over 200 online hardship requests from 2017 to 2023 [s71]. The customers had asked for breathing room because of domestic abuse, natural disasters, serious illness or job loss [s71]. Some were left with adverse credit records and had their debts sold to buyers who pursued them [s71]. The judge said the contraventions were not suggested to be deliberate and arose instead from "inadequate systems and operational failures", but accepted that they were "grossly negligent" [s71].

The controls that matter most. Prioritise: let the model plan contact, not start consequential steps. Risk-assess: identify vulnerable borrowers before any consequential step, such as default reporting, a debt sale or a court order like the one in Tarihal. Orchestrate: a hardship notice stops automated chasing until a person who can grant time or a lower payment has replied. Validate: follow up borrowers who stopped answering, not only those who paid. Embed: report hardship response times next to cash collected.

The paired measures. Cash collected, set against hardship requests answered on time, complaints, re-default within 12 months and arrangements that hold.

The owner. The head of collections as business owner, with a named risk owner for vulnerable customers who can stop an action.

Trading: when the model moves the market

What it is rewarded for. Profit, measured against risk limits.

How it shapes its own evidence. A strategy that trades enough moves the prices it is trying to predict, and the profit it books can be partly a record of that impact. When many firms follow similar signals, their exits can become the event.

What happened. In early April 2025, a "more orderly unwinding" of relative-value trades tied to interest rate swap markets "seems to have contributed to the heightened volatility" in US Treasury markets, the BIS said in June 2025 [s72]. The Bank of England's Financial Policy Committee noted in April 2025 that funds had so far met significant margin calls without actions that would "further amplify" the volatility [s73]. The rest of the evidence looks forward. The Bank's July 2026 Financial Stability Report says the evidence to date suggests trading firms are using more autonomous AI mainly for research, coding and surveillance, rather than fully autonomous trading [s74]. If AI systems start to outperform in testing, it warns, firms may rely on them more directly in ways that "increase the risk of correlated behaviour" [s74]. The BIS said in June 2026 that supervisors need better visibility to capture "model-driven correlated exposures, herding behaviour and the potential for algorithmic collusion" [s75].

The controls that matter most. Prioritise: keep models at research and signals, with people placing orders. Risk-assess: estimate how far the strategy's own trades move the prices it books. Orchestrate: set position, leverage and order-rate limits in the trading system rather than the model, and give a person the authority to stop it. Validate: stress the strategy against its own exit, and against the same exit by others holding similar positions. Embed: review every use of that stop with market risk.

The paired measures. Profit and loss, read with market impact, crowding with similar strategies, and liquidity under stress.

The owner. The desk head as business owner, though market risk can cut limits without the desk's agreement.

What the five loops hide

Loop	Rewarded for	Becomes unobservable	Paired measures	Owner
Credit	Repaid loans	Declined applicants; loans from other lenders	Defaults with affordability and multiple borrowing	Business owner, head of credit policy
Fraud	Fraud stopped	Genuine payments blocked	Losses with false blocks and time to restore	Business owner, head of fraud
AML	Confirmed alerts	Activity never monitored	Alert precision with coverage and blind samples	Compliance owner, MLRO
Collections	Cash recovered	Hardship and re-default	Cash with hardship, complaints, re-default	Business owner, head of collections
Trading	Profit	Market impact and crowding	P&L with impact and stress liquidity	Business owner, desk head

Our summary of the five loops.

Every loop rewards a number that the system's own decisions help produce, from low defaults to profit, and hides a second number behind it. The PROVE questions are a way of finding that second number before the system is given authority, rather than after a regulator, a newspaper or a market does.

Once the evidence is in, someone has to decide how much authority the system gets, on what conditions, and who answers for it.

6. The deployment decision

Boards are often asked about AI with too little in front of them. In PwC's September 2026 survey of US corporate directors, 82% rated the information linking AI outcomes, risks and business performance as fair, poor or not provided [s76]. A committee in that position is usually handed a yes-or-no question. The better question is the one Section 5 ended on, with one addition: for how long.

Three answers, not two

The useful answers are three:

- **1. Deploy at a stated step.** The system may answer, recommend, decide or execute, as Section 3 set out, and the paper says which. The limits travel with the permission.
- **2. Deploy narrower.** The evidence supports less than was asked for: recommend rather than decide, a smaller customer group, a lower transaction limit, a person releasing every block.
- **3. Wait.** A stage of PROVE has a gap that narrowing cannot close, and the paper says what evidence would change the answer.



Figure 22: Three answers a committee can give. Each one asks something different of the evidence.

Each answer asks something different of the evidence. Deploying at the requested step needs every stage covered for that step, not for the one below it. Deploying narrower needs the evidence to support the smaller permission on its own terms, so that the narrow version is a real system rather than the big one with a hopeful label. Waiting needs a named gap, an owner for the missing work and a date to look again, or it quietly becomes a no.

Whichever answer the committee gives, it has to exist in the software as well as in the minutes. If the committee approves recommendation only, the system's credentials must lack the power to turn a recommendation into a credit decision. Otherwise the minutes and the system describe different businesses.

What the decision record holds

A board paper that grants authority should answer six things in writing:

- **Evidence by stage.** For each PROVE stage, what was shown and what is still missing, with the gaps marked rather than smoothed over.

- **The permitted authority.** Which of the four steps, for which products and customers, inside which limits, enforced by which system, and for which model and version.
- **Unresolved uncertainty.** What the tests could not show, and how that narrows the permission. A disappointing result and a missing observation are different things: an applicant who was turned away and never heard from again is not a correct rejection. Where the missing rate cannot be estimated, the paper names which of the three choices in Section 4 the firm made.
- **Named owners.** The four owners from Embed, each with a deputy, and someone a customer can appeal to.
- **Intervention triggers.** The events that narrow or suspend the permission without waiting for the next committee.
- **A dated review.** The day the permission ends unless it is renewed with new evidence.



Figure 23: Six things a board paper that grants authority should answer in writing.

Many firms have not yet written this down. In a cross-sector Kyndryl survey published in November 2026, 64% of leaders reported positive returns on AI, but only 27% had formally updated roles and responsibilities around it [s77]. The FCA's October 2026 review of AI in financial services shows what good practice looks like: one firm mapped every material AI use case to a named senior manager, with documented risk appetite and escalation routes before deployment [s35]. Firms building agents registered them in a central inventory, which the FCA said can help a firm see which are live, who owns them, what systems and data they can reach and what controls apply [s35]. A small firm recorded whether staff needed to review AI outputs, whether they did, and whether they approved, overrode or escalated them, so the record showed oversight working rather than assumed [s35].

Triggers and expiry

A permission that never expires becomes a habit. The record should say in advance what takes authority away:

- a paired measure from Section 5 moving the wrong way, such as wrongful blocks rising while fraud losses fall;
- drift in the inputs or in the population the system sees;
- any change Section 4 sends back through approval, such as a vendor model update or a request for wider permission;
- a spike in complaints, appeals or ombudsman cases;

- the follow-up evidence itself stopping, so that a quiet dashboard means nothing is being measured;
- an outside warning, from law enforcement, a regulator or the press;
- an incident, or a test the system fails on rerun.

Trigger	Who acts	Default action
A paired measure moves the wrong way	Business owner	Narrow
Drift in inputs or population	Technical owner	Narrow or suspend
A change Section 4 sends back through approval	Technical owner	Back through approval
Spike in complaints, appeals or ombudsman cases	Business owner	Suspend and sample
Follow-up evidence stops arriving	Risk owner	Narrow to the last step evidenced
Outside warning	Compliance owner	Review at once
Incident or failed rerun	Risk owner	Roll back

An illustration of the record, not a regulatory list.

Vendor updates belong on the list because they happen without the firm touching its own code. On 25 April 2025 OpenAI rolled out an update to GPT-4o in ChatGPT that, in the company's words, made the model "noticeably more sycophantic" [s78]. Four days later it said it had rolled the update back, so people were using an earlier version "with more balanced behavior" [s79]. That happened in a consumer product, but a bank's system built on a vendor's model can change underneath a permission in the same way, without anyone approving the change. FINRA's January 2026 oversight report lists tracking "which model version was used and when" among the practices firms may adopt [s80].

Outside warnings belong because a system trained on internal history cannot see them. In December 2025 the FCA fined Barclays £39.3 million over its handling of Stunt & Co, a customer that received £46.8 million from a money laundering operation [s81]. The bank did not properly consider the risk "even after receiving information from law enforcement" and after learning of police raids, and reviewed its exposure only once it learned that NatWest was to be prosecuted [s81]. No AI was involved. The warning came from outside the bank's own data, and the bank did not act on it.

Live use can also overturn the reason for the permission. In July 2025 Commonwealth Bank of Australia announced 45 job cuts as it introduced an AI voice bot, then reversed the decision the following month [s82], saying the roles "were not redundant" [s83]. A case built on a forecast should name the live measure that would show the forecast was wrong.

Each trigger needs an owner who acts on it, a default action (narrow, suspend or roll back) and a fallback that carries the work when the system stops. MAS's October 2026 guidelines ask firms with high-risk AI for contingency plans with fallback options, "such as alternative systems or manual processes" [s49], and to consider "kill switches" or override mechanisms that can "rapidly deactivate the AI system if it exceeds risk tolerances" [s49]. The Hong Kong Monetary Authority issued its guidance on the responsible adoption of AI on 15 October 2026 [s84]. Its annex says more material or higher-risk uses should have senior management sign-off, independent validation and "a kill switch or equivalent", with records that allow outcomes to be traced, according to a regulatory publisher's summary [s85]. Once the EU AI Act's high-risk duties reach credit scoring, Article 26 will require a deployer that has reason to consider a system may present a risk to inform the provider or distributor and the market surveillance authority "without undue delay", "and shall suspend the use of that system" [s86]. For banks, the same

article treats the duty to monitor the system as met by complying with their existing financial-services governance rules [s86], so the switch belongs in the governance the bank already runs. A trigger without a working switch is a wish.

The review date matters as much. A permission granted for a fixed term, renewable only on fresh evidence from cases the system did not choose, keeps the feedback loop from quietly becoming the evidence. Renewal is for whoever granted the permission to decide, on the risk owner's advice, and not for a vendor's release schedule.

A worked example: Meena's loan, at scale

Section 1's Try it box asked you to decide one loan, for Meena, who runs a home bakery, has 14 months of credit history and earns most of her year at festival time. Here is the same decision made as a permission. This is an illustration: the lender, its people, its dates, its figures and its limits are made up.

A lender wants a model to decide small-business working-capital loans under £25,000 (illustrative, not calculated from a source), and to pay them out without a person signing each one. The paper would read something like this.

- **Evidence by stage.**
- Prioritise: a baseline from 12 months of human decisions, with defaults, arrears and wrongful refusals sampled.
- Risk-assess: thin files and irregular income named as the main exclusion risk, and borrowing from other lenders at the same time as the main harm risk.
- Orchestrate: a payout limit enforced in the payments system, a referral rule for incomplete records, and an appeal route staffed by underwriters.
- Validate: a loss-capped sample of borderline applicants approved for comparison, and tests on seasonal incomes like Meena's.
- Embed: owners named, monitoring live.
- The gap: no evidence yet on applicants' total debt elsewhere.
- **Permitted authority.** Recommend for applicants with less than 24 months of credit history. Execute for the rest up to £10,000 (illustrative, not calculated from a source): the model decides and pays out. Decide above that: the model decides, but an underwriter signs every payout. The model may not decline an incomplete file.
- **Uncertainty.** The gap on other borrowing keeps the limit low and keeps the thin-file group at Recommend.
- **Owners.** Priya Nair, head of small-business credit, for the business; Tom Okafor, technical owner, for the model and its rollback; Lena Fischer, credit risk officer, who can suspend it; Wei Ling Tan, compliance lead, for notices and records; and Sam Reid, who handles appeals. Each has a named deputy.
- **Triggers.** Arrears in the decided group above the human baseline for two months running; upheld appeals above a set share; the comparison sample stopping; any change to the vendor's scoring model.
- **Review.** The permission starts on 1 December 2026, is reviewed on 1 June 2027 and lapses on 30 November 2027 unless renewed on new evidence from the comparison sample.

Under this paper the model would not decide Meena's loan at all. With 14 months of history she is in the Recommend group, so an underwriter would make the call the Try it box gave you, with her score of 0.58 against the 0.60 cut-off, and the model's uncertainty, in front of them. An applicant with three years of credit history would have the loan decided and paid out by the model up to £10,000 (illustrative, not calculated from a source). Meena's group moves up a step only when the comparison sample shows how borrowers like her actually repay. If the underwriter turns her down, she is owed reasons. In the United States, Regulation B already requires a lender's statement of reasons for adverse action to be "specific" and to "indicate the principal reason(s)" [s87].



Figure 24: The permission for Meena's loan. An illustration. The lender, the limits and the dates are made up.

The paper does not say the model is good. It says what the model has earned the right to do, for whom, and what would take that right away.

The regulator as circuit breaker

India shows why the institution needs its own brakes. On 16 November 2023, after flagging high growth in consumer credit, the RBI raised risk weights on specified bank consumer-credit exposures by 25 percentage points to 125%, with exclusions including housing, education, vehicle and gold-secured loans [s90]. A separate 25-point increase on certain bank loans to non-bank lenders, made in the same circular, was reversed from 1 April 2025 [s91].

Default-loss guarantees, commonly called FLDG, compensate a lender for part of its loss when a borrower defaults [s92]. Under its 8 June 2023 rules, the RBI capped such cover at 5% of the specified loan portfolio, as a published legal commentary explains [s92]. That circular has since been repealed, and the RBI points readers to its Digital Lending Directions of 8 May 2025 [s93]. The cap limits the guarantee; it is not a finding that the underlying loans are affordable.

In August 2024, the RBI announced a move from monthly to fortnightly credit reporting [s94]. Faster reporting shortens the gap in which new borrowing may be invisible; it does not make a bureau report a live view of every debt.

The three measures address different weaknesses: the capital behind lending, reliance on a partner's guarantee, and the lag in seeing a borrower's obligations. Nor did the RBI wait for losses: it asked lenders to address any build-up of risk [s90]. The point for PROVE is who acts first. A supervisor's tightening belongs on the trigger list as an outside warning, not in place of limits the lender has set and tested itself. Proving only after deployment is too late for the borrower who needed that limit to hold.

What has to be true

Section 3 set out what would weaken our hypothesis. Little in the public record does. Firms report written policies far more often than they report watching outcomes after launch, and the harms in Section 3 came to light through a journalist, a state attorney general and an ombudsman rather than through the firms' own checks. That does not prove the hypothesis, because the record cannot tell a hidden harm from an absent one. It does put the burden of proof on the system.

In finance, decisions help shape the evidence used to judge them. An AI system given authority joins that loop. It does not have to be malicious or broken to mislead the people relying on it. It only has to be judged on outcomes it helped to choose.

The answer is not to keep AI out of finance, and it is not to trust a benchmark. It is to grant authority the way a careful institution grants any authority: in steps, against evidence, with limits it can enforce and a named person who answers when it goes wrong. Prove the value, the risks, the controls, the performance and who owns the outcome. Then deploy, and keep proving it.

Sources

Source numbers match the web edition. Limits are noted after each source.

1. What is money? Bank of England explainer, 27 February 2025 (regulator). https://www.bankofengland.co.uk/explainers/what-is-money

2. BNB Chain Ecosystem Update. BNB Chain Blog, 7 October 2022 (primary). https://www.bnbchain.org/en/blog/bnb-chain-ecosystem-update

3. $570 million worth of Binance's BNB token stolen in another major crypto hack [s3]. CNBC, 7 October 2022 (press). https://www.cnbc.com/2022/10/07/more-than-100-million-worth-of-binances-bnb-token-stolen-in-another-major-crypto-hack.html *Dollar value is at prices on the day of the report.*

4. Karnataka: Woman, newborn pushed out of their house for not paying loan instalments. The New Indian Express, 25 January 2025 (press). https://www.newindianexpress.com/states/karnataka/2025/Jan/25/karnataka-woman-newborn-pushed-out-of-their-house-for-not-paying-loan-instalments *A press report of one family's case; it does not say how the loan was underwritten.*

5. To Score Is to Decide: About the SCHUFA case. Verfassungsblog, 14 December 2023 (legal commentary). https://verfassungsblog.de/to-score-is-to-decide/

6. CJEU rules that a credit score constitutes automated decision making under the GDPR. A&O Shearman, 11 January 2024 (law-firm briefing). https://www.aoshearman.com/en/insights/ao-shearman-in-the-netherlands/102k8l4/cjeu-rules-that-a-credit-score-constitutes-automated-decision-making-under-the-gd

7. The Illusion of Improvement: Reject Inference Strategies in Credit Scoring. arXiv preprint 2606.18479 (June 2026). https://arxiv.org/html/2606.18479v1 *Simulation over ten retraining cycles on public benchmark credit data; figures quoted are for the Default dataset with gradient-boosted trees, averaged over five runs. Not peer reviewed; not a bank portfolio.*

8. How Flawed Data Aggravates Inequality in Credit. Stanford HAI, 6 August 2021 (press summary of Blattner and Nelson). https://hai.stanford.edu/news/how-flawed-data-aggravates-inequality-credit *US credit data only.*

9. How Costly is Noise? Data and Disparities in Consumer Credit. Laura Blattner and Scott Nelson, arXiv 2105.07554 (May 2021). https://arxiv.org/abs/2105.07554 *Working paper; the halving estimate comes from a structural model of the US mortgage market.*

10. Performative Prediction. Perdomo, Zrnic, Mendler-Dünner and Hardt, arXiv 2002.06673 (2020). https://arxiv.org/abs/2002.06673

11. Reject inference methods in credit scoring. Ehrhardt et al., via PMC, 2021. https://pmc.ncbi.nlm.nih.gov/articles/PMC9041715/

12. Feedback Loops in ML Systems Guide. AI Understanding, 2 October 2026 (explainer). https://aiunderstanding.org/nl/learn/feedback-loops-in-ml-systems *Secondary explainer, used only for a practice recommendation.*

13. How AI assistance impacts the formation of coding skills. Anthropic research, 2026 (randomised controlled trial). https://www.anthropic.com/research/AI-assistance-coding-skills *52 mostly junior engineers learning one new library; the quiz was taken minutes after the task.*

14. We are Changing our Developer Productivity Experiment Design. METR, 24 February 2026 (research update). https://metr.org/blog/2026-02-24-uplift-update/ *METR says its late-2025 data is an unreliable signal; both newer intervals include zero.*

15. Navigating the Jagged Technological Frontier. Dell'Acqua et al., Organization Science, 11 March 2026 (peer-reviewed paper). https://pubsonline.informs.org/doi/10.1287/orsc.2025.21838 *Field experiment run with GPT-4.*

16. Back to the Beginnings of AI at Work. Harvard Business School AI Institute, 20 April 2026 (summary of the Harvard and BCG field experiment). https://aiinstitute.hbs.edu/back-to-the-beginnings-of-ai-at-work/ *A summary page, not the paper itself.*

17. AA-Omniscience: Evaluating Cross-Domain Knowledge Reliability in Large Language Models. Artificial Analysis, arXiv 2511.13029 (November 2025). https://arxiv.org/html/2511.13029v1 *Defines the benchmark and its metrics.*

18. AA-Omniscience Accuracy Leaderboard. BenchLM, table updated 11 November 2026. https://benchlm.ai/benchmarks/omniscienceaccuracy *Leaderboards change; figures are as of the date shown.*

19. AA-Omniscience Hallucination Rate Leaderboard. BenchLM, table updated 11 November 2026. https://benchlm.ai/benchmarks/omnisciencehallucinationrate *Every row on this page was carried from another source; none was run independently.*

20. 'AI disclaimers not enough to protect consumers', FCA warned. FT Adviser, 19 October 2026 (press, reporting research by Saturn). https://www.ftadviser.com/content/16bfe791-b923-4742-8ac2-2ced2b030c22 *Press report of a vendor study; the method is not published here.*

21. τ-Knowledge: Evaluating Conversational Agents over Unstructured Knowledge. arXiv 2603.04370 (March 2026). https://arxiv.org/abs/2603.04370 *Controlled benchmark tasks with models of early 2026; not live customer outcomes.*

22. Artificial intelligence in UK financial services - 2026. Bank of England and FCA, 26 October 2026. https://www.bankofengland.co.uk/report/2026/artificial-intelligence-in-uk-financial-services-2026 *Survey of UK financial firms; self-reported by respondents.*

23. LLM07:2025 System Prompt Leakage. OWASP GenAI Security Project, 2025. https://genai.owasp.org/llmrisk/llm072025-system-prompt-leakage/ *Design guidance, not a measured failure rate.*

24. Mistaking AI vulnerability could lead to large-scale breaches, NCSC warns. UK National Cyber Security Centre, 10 December 2025. https://www.ncsc.gov.uk/news/mistaking-ai-vulnerability-could-lead-to-large-scale-breaches *Security guidance, not a count of breaches.*

25. Zombies With Wallets: The Dark Side of AI Payments (State of AI Agent Payments 2026). BeInCrypto, 3 November 2026. https://beincrypto.com/ai-zombies-wallets-consciousness-agent-payments/ *Crypto industry report; the Bankrbot transfer was a linked-wallet exploit, not a regulated bank loss.*

26. Scoop: Unvetted AI agents are accessing U.S. bank accounts. Axios, 22 October 2026. https://www.axios.com/2026/10/22/ai-agents-banks-unauthorized-accounts-access *Press report based on unnamed sources; not supervisory data.*

27. Massachusetts AG Settles Fair Lending Action Based Upon AI Underwriting Model. Consumer Finance & Fintech Blog (cfsreview.com), 24 July 2025. https://www.cfsreview.com/2025/07/massachusetts-ag-settles-fair-lending-action-based-upon-ai-underwriting-model/ *Law-firm commentary; used for the announcement date (10 July 2025).*

28. AG Campbell Announces $2.5 Million Settlement With Student Loan Lender For Unlawful Practices Through AI Use [s28]. Office of the Massachusetts Attorney General, 10 July 2025. https://www.mass.gov/news/ag-campbell-announces-25-million-settlement-with-student-loan-lender-for-unlawful-practices-through-ai-use-other-consumer-protection-violations *Allegations resolved by settlement (assurance of discontinuance); not findings of a court.*

29. Our house purchase is at risk from a frozen Revolut account. The Guardian (consumer column), 4 November 2026. https://www.theguardian.com/money/2026/nov/04/our-house-purchase-is-at-risk-from-a-frozen-revolut-account *One reader's case as reported; does not say whether a model or a person triggered the freeze.*

30. Money mules: mule activity and cashing out findings. Financial Conduct Authority, 23 September 2026. https://www.fca.org.uk/publications/multi-firm-reviews/money-mules-activity-cashing-out-findings *Counts closures of suspected mule accounts; it does not measure wrongful closures.*

31. Final decision DRN-5687237 (Revolut Ltd, AI Chat Assistant). Financial Ombudsman Service, August 2025. https://www.financial-ombudsman.org.uk/decision/DRN-5687237.pdf *One complaint decided by the ombudsman; not a regulatory finding.*

32. Ombudsman decisions record Revolut's AI chat assistant giving customers false payment and refund information. Banking News AI (incident record), 21 October 2026. https://www.bankingnewsai.com/ai-incidents/revolut-2025-ai-chat-assistant-ombudsman-errors *Press incident database summarising the two decisions; cited for dates and Revolut's terms.*

33. Final decision DRN-6062372 (Revolut Ltd, chatbot refund). Financial Ombudsman Service, January 2026. https://www.financial-ombudsman.org.uk/decision/DRN-6062372.pdf *One complaint decided by the ombudsman; not a regulatory finding.*

34. UK consumers warned over AI chatbots giving inaccurate financial advice. The Guardian, reporting Which? tests, 18 November 2025. https://www.theguardian.com/technology/2025/nov/18/warning-ai-chatbots-inaccurate-financial-advice-tips-chatgpt-copilot-uk *Consumer-group tests of general-purpose chatbots, not a bank's own assistant; no customer loss reported.*

35. AI in financial services: good practice and areas for improvement. Financial Conduct Authority, 20 October 2026. https://www.fca.org.uk/publications/good-poor-practice/ai-financial-services *Supervisory observations across firms; not a rule or an enforcement finding.*

36. Bias in the Loop: How Humans Evaluate AI-Generated Suggestions. Harvard Data Science Review, 30 April 2026. https://hdsr.mitpress.mit.edu/pub/nrcn4h7d *Online experiment checking emissions tables, not bank underwriting.*

37. FSB finalises sound practices for the responsible adoption of artificial intelligence (AI). Financial Stability Board, 15 October 2026. https://www.fsb.org/2026/10/fsb-finalises-sound-practices-for-the-responsible-adoption-of-artificial-intelligence-ai/ *Press release for the final report; sound practices, not binding rules.*

38. SR 26-2: Revised Guidance on Model Risk Management. Board of Governors of the Federal Reserve System, 17 April 2026. https://www.federalreserve.gov/supervisionreg/srletters/SR2602.htm *Supervisory guidance; non-binding by its own terms.*

39. Model Risk Management: Revised Guidance (Bulletin 2026-13). Office of the Comptroller of the Currency, 17 April 2026. https://www.occ.gov/news-issuances/bulletins/2026/bulletin-2026-13.html *Supervisory guidance for US banks; generative and agentic AI out of scope.*

40. AI Governance and Model Risk Platforms: Buyer's Guide. Finantrix, 5 November 2026. https://www.finantrix.com/buyer-guides/ai-governance-model-risk-platforms *A vendor buyer's guide; cited only for its quotation of the guidance's scope.*

41. Agencies Revise the Interagency Model Risk Management Guidance (FIL-15-2026). Federal Deposit Insurance Corporation, 17 April 2026. https://www.fdic.gov/news/financial-institution-letters/2026/agencies-revise-interagency-model-risk-management-guidance *Summary of the interagency guidance.*

42. MAS Sets Out Supervisory Expectations on Responsible AI Adoption by Financial Institutions. Monetary Authority of Singapore, 7 October 2026. https://www.mas.gov.sg/news/media-releases/2026/mas-sets-out-supervisory-expectations-on-responsible-ai-adoption-by-financial-institutions *Media release for the Guidelines; effective from 7 October 2027.*

43. EU AI Act obligations by role and date, article by article. EON Risk Services, 2 November 2026. https://www.eonriskservices.com/eu-ai-act-obligations *A compliance consultancy's summary of the Act after the Digital Omnibus.*

44. AI experiences rapid adoption, but with mixed outcomes: highlights from VotE: AI & Machine Learning. S&P Global Market Intelligence, 30 May 2025. https://www.spglobal.com/market-intelligence/en/news-insights/research/ai-experiences-rapid-adoption-but-with-mixed-outcomes-highlights-from-vote-ai-machine-learning *Cross-sector self-reported survey of 1,006 professionals in North America and Europe, not banks alone.*

45. Klarna AI assistant handles two-thirds of customer service chats in its first month. Klarna (press release), 27 February 2024. https://www.klarna.com/international/press/klarna-ai-assistant-handles-two-thirds-of-customer-service-chats-in-its-first-month/ *The company's own launch claims, not an independent measure.*

46. Klarna CEO Reverses Course By Hiring More Humans, Not AI. Entrepreneur, reporting Bloomberg, May 2025 (page updated January 2026). https://www.entrepreneur.com/business-news/klarna-ceo-reverses-course-by-hiring-more-humans-not-ai/491396 *Press summary of a Bloomberg interview and Klarna's 2024 claims.*

47. Klarna Rethinks AI Cost-Cutting Plan With Call for Real People. Bloomberg, 8 May 2025. https://www.bloomberg.com/news/articles/2025-05-08/klarna-turns-from-ai-to-real-person-customer-service *Press interview; the full article is behind a paywall.*

48. SEC Charges Restaurant-Technology Company Presto Automation for Misleading Statements About AI Product. US Securities and Exchange Commission, 14 January 2025. https://www.sec.gov/enforcement-litigation/administrative-proceedings/33-11352-s *Settled order outside banking; Presto neither admitted nor denied the findings.*

49. Guidelines on Artificial Intelligence Risk Management. Monetary Authority of Singapore, 7 October 2026. https://www.mas.gov.sg/regulation/guidelines/guidelines-on-artificial-intelligence-risk-management-for-financial-institutions *The Guidelines' own text; sections 3 to 4 apply from 7 October 2027 and sections 5 to 6 from 7 October 2028.*

50. Singapore issues new AI Risk Management Guidelines for financial institutions. Herbert Smith Freehills Kramer (hlc.com), 30 October 2026. https://www.hlc.com/en/publications/mas-issues-new-ai-risk-management-guidelines-for-financial-institutions *A law firm's summary of the MAS Guidelines, not the Guidelines' own text.*

51. 'Rogue' AI agent went haywire at tech company. The CEO is still 'bullish' on the technology. ABC News, 30 April 2026. https://abcnews.com/GMA/News/rogue-ai-agent-haywire-tech-company-ceo-bullish/story?id=132473181 *Press report of a software company incident outside regulated finance; the data was restored.*

52. Singapore's MAS sharpens oversight as AI risks mount. QA Financial, 13 October 2026. https://qa-financial.com/mas-sharpens-testing-expectations-as-ai-risks-swell/ *Trade press; SAFR is an industry reference framework, not a supervisory rule.*

53. CFPB Orders Operator of Cash App to Pay $175 Million and Fix Its Failures on Fraud [s53]. Consumer Financial Protection Bureau, January 2025 (archived page). https://www.consumerfinance.gov/archive/newsroom/cfpb-orders-operator-of-cash-app-to-pay-175-million-and-fix-its-failures-on-fraud/ *Enforcement order on earlier conduct; does not establish an AI cause.*

54. SS1/23: Model risk management principles for banks (April 2026 version). Bank of England, Prudential Regulation Authority, 16 April 2026. https://www.bankofengland.co.uk/-/media/boe/files/prudential-regulation/supervisory-statement/2026/liaf0126app5.pdf *The supervisory statement's own text; applies to firms with internal model approval for regulatory capital.*

55. Enterprise AI Enters an ROI Reckoning: More Use, Uneven Value, Reveals Deloitte Survey. Deloitte, 29 October 2026. https://www.deloitte.com/us/en/about/press-room/enterprise-ai-enters-an-roi-reckoning.html *Cross-industry self-reported survey (May to June 2026), not bank-only.*

56. FCA fines Monzo £21m for failings in financial crime controls [s56]. Financial Conduct Authority, 8 July 2025. https://www.fca.org.uk/news/press-releases/fca-fines-monzo-21m-failings-financial-crime-controls *Enforcement on controls from 2018 to 2022; not an AI case.*

57. Consumer & Community Context: Alternative Data: Expanding Access to Credit. Board of Governors of the Federal Reserve System, 22 October 2025. https://www.federalreserve.gov/publications/2025-october-consumer-community-context.htm *US estimates and survey data; not a measure of any lender's model.*

58. FinRegLab Study Shows Cash-Flow Data Can Expand Small Business Lending. FinRegLab, 3 June 2025. https://finreglab.org/press-releases/finreglab-study-shows-cash-flow-data-can-expand-small-business-lending/ *Two fintech lenders' small-business loans; a research nonprofit's release.*

59. CFPB Research Reveals Heavy Buy Now, Pay Later Use Among Borrowers with High Credit Balances and Multiple Pay-in-Four Loans. Consumer Financial Protection Bureau, January 2025 (archived page). https://www.consumerfinance.gov/archive/newsroom/cfpb-research-reveals-heavy-buy-now-pay-later-use-among-borrowers-with-high-credit-balances-and-multiple-pay-in-four-loans/ *2021 to 2022 data from six large BNPL firms matched to credit records.*

60. Why Asia's biggest microfinance lenders face debt stress. The Banker, 20 August 2026. https://www.thebanker.com/content/b6d494f2-1a28-4492-850a-55081cd9562c *Press analysis quoting an S&P Global Ratings analyst and ASA.*

61. Rating Rationale: Fusion Finance Ltd. CRISIL Ratings, 30 December 2025. https://www.crisilratings.com/mnt/winshare/Ratings/RatingList/RatingDocs/FusionFinanceLimited_December%2030_%202025_RR_385412.html *A rating rationale for one lender; it does not attribute outcomes to AI or to any model.*

62. FCA regulation of deferred payment credit: what BNPL firms must do before July 2026. PKF Littlejohn, 9 June 2026. https://www.pkf-l.com/insights/fca-bnpl-regulation-2026-rules-timeline-compliance/ *An accounting firm's summary of the FCA regime.*

63. Why stronger Buy Now Pay Later protections are the right step for consumers. Financial Conduct Authority (blog), 26 June 2026. https://www.fca.org.uk/news/blogs/why-stronger-buy-now-pay-later-protections-are-right-step-consumers *Regulator's blog, not the rules themselves.*

64. Joint EBA-ECB report on payment fraud: strong authentication remains effective but fraudsters are adapting. European Central Bank and European Banking Authority, 15 December 2025. https://www.ecb.europa.eu/press/pr/date/2025/html/ecb.pr251215~e133d9d683.en.html *Industry-reported fraud across the EEA, 2022 to 2024.*

65. APP Scams Reimbursement Dashboard. Payment Systems Regulator, updated 30 July 2026. https://www.psr.org.uk/information-for-consumers/app-scams-reimbursement-dashboard/ *Closed Faster Payments claims reported by sending firms; not validated against receiving firms.*

66. Payment fraud falls by £73m following PSR reimbursement scheme [s66]. Payment Systems Regulator, 1 July 2026. https://www.psr.org.uk/news-and-updates/latest-news/news/payment-fraud-falls-by-73m-following-psr-reimbursement-scheme/ *The regulator's announcement of the independent Frontier Economics evaluation.*

67. APP fraud reimbursement requirement: were industry's concerns justified? Grant Thornton, 21 September 2026. https://www.grantthornton.co.uk/insights/app-fraud-reimbursement-requirement-were-industrys-concerns-justified/ *Advisory firm commentary on the evaluation.*

68. United States of America v. TD Bank, N.A. US Department of Justice, October 2024. https://www.justice.gov/criminal/case/united-states-america-v-td-bank-na *Plea agreement facts; conduct from 2014 to 2024; not an AI case.*

69. FCA fines Nationwide £44m for failings in financial crime controls [s69]. Financial Conduct Authority, 12 December 2025. https://www.fca.org.uk/news/press-releases/fca-fines-nationwide-44m-failings-financial-crime-controls *Enforcement on controls from 2016 to 2021; not an AI case.*

70. Can AI Replace Human Debt Collectors? Yale School of Management Insights, 20 May 2025. https://insights.som.yale.edu/insights/can-ai-replace-human-debt-collectors *Summary of a working paper using 2021 to 2022 cases in China; repayment, not borrower welfare.*

71. Federal Court orders Westpac to pay $26 million penalty for hardship failures [s71]. Australian Securities and Investments Commission, 27 May 2026. https://www.asic.gov.au/about-asic/news-centre/find-a-media-release/2026-releases/26-107mr-federal-court-orders-westpac-to-pay-26-million-penalty-for-hardship-failures *Court penalty for system failures from 2017 to 2023; not an AI case.*

72. Annual Economic Report 2025, chapter I: Sustaining stability amid uncertainty and fragmentation. Bank for International Settlements, 29 June 2025. https://www.bis.org/publications/aer-2025/sustaining-stability-amid-uncertainty-fragmentation *Cautious causal wording; not an AI episode.*

73. Financial Policy Committee Record, April 2025. Bank of England, 9 April 2025. https://www.bankofengland.co.uk/financial-policy-committee-record/2025/april-2025 *A record of the committee's assessment at the time.*

74. Financial Stability Report, July 2026. Bank of England, 7 July 2026. https://www.bankofengland.co.uk/financial-stability-report/2026/july-2026 *Assessment and warning, not measured AI trading harm.*

75. Annual Economic Report 2026, chapter I: Progress and peril. Bank for International Settlements, 28 June 2026. https://www.bis.org/publications/aer-2026/progress-peril *Policy recommendation.*

76. PwC's 2026 Annual Corporate Directors Survey. PwC US Governance Insights Center, 23 September 2026. https://www.pwc.com/us/en/about-us/newsroom/press-releases/pwc-2026-annual-corporate-directors-survey.html *Survey of US public-company directors across sectors, not bank boards alone; self-reported.*

77. Kyndryl Report: AI returns rise as enterprise readiness falls. Kyndryl, 10 November 2026. https://kyndryl.gcs-web.com/news-releases/news-release-details/kyndryl-report-ai-returns-rise-enterprise-readiness-falls *Cross-sector survey commissioned by an IT infrastructure provider; self-reported.*

78. Expanding on what we missed with sycophancy. OpenAI, 2 May 2025. https://openai.com/index/expanding-on-sycophancy/ *The vendor's own account of a consumer product; passage taken from the search result.*

79. Sycophancy in GPT-4o: What happened and what we're doing about it. OpenAI, 29 April 2025. https://openai.com/index/sycophancy-in-gpt-4o/ *The vendor's own account; passage taken from the search result.*

80. 2026 FINRA Annual Regulatory Oversight Report: GenAI. Financial Industry Regulatory Authority, January 2026. https://www.finra.org/rules-guidance/guidance/reports/2026-finra-annual-regulatory-oversight-report/gen-ai *For US broker-dealers; examination considerations, not new rules.*

81. FCA fines Barclays £42 million for poor handling of financial crime risks [s81]. Financial Conduct Authority, 5 December 2025. https://www.fca.org.uk/news/press-releases/fca-fines-barclays-42-million-poor-handling-financial-crime-risks *Enforcement on legacy monitoring controls; not an AI case.*

82. Commonwealth Bank backtracks on AI job cuts, apologises. ABC News, 21 August 2025. https://www.abc.net.au/news/2025-08-21/cba-backtracks-on-ai-job-cuts-as-chatbot-lifts-call-volumes/105679492 *Press report; passage taken from the search result.*

83. CBA reverses AI-driven job cuts, admits 'error'. Information Age (ACS), 21 August 2025. https://ia.acs.org.au/article/2025/cba-reverses-ai-driven-job-cuts--admits--error-.html *Press report; passage taken from the search result.*

84. Responsible Adoption of Artificial Intelligence. Hong Kong Monetary Authority, 15 October 2026. https://brdr.hkma.gov.hk/eng/doc-ldg/current/20261015-4-EN *Register entry for the circular; the circular text itself not read.*

85. HKMA ART Framework: Nine AI Principles for Hong Kong Banks. RegReportingDesk (regulatory publisher), 19 October 2026. https://regreportingdesk.com/hkma-art-framework-responsible-ai-guidance/ *A publisher's summary of the HKMA annex, not the annex.*

86. Article 26: Obligations of deployers of high-risk AI systems. European Commission, AI Act Service Desk (text of Regulation (EU) 2024/1689), read 13 November 2026. https://ai-act-service-desk.ec.europa.eu/en/ai-act/article-26 *The Regulation's own text; applies to credit scoring once the high-risk duties apply.*

87. 12 CFR § 1002.9 Notifications (Regulation B). Consumer Financial Protection Bureau, 1 January 2025. https://www.consumerfinance.gov/rules-policy/regulations/1002/9 *US rule on credit notifications; does not govern model validation.*

88. Global Finance & Technology Network: Who We Are. Global Finance & Technology Network (GFTN), 2026. https://gftn.co/who-we-are/ *The publisher's own institutional description on gftn.co and gftn.co/who-we-are/.*

89. Digital Credit Algorithms in India: Expanding Access or Cementing Exclusion? Center for Financial Inclusion, 2022; page metadata 17 April 2024. https://www.centerforfinancialinclusion.org/article/digital-credit-algorithms-in-india-expanding-access-or-cementing-exclusion/ *Historical qualitative research on Indian migrants; not a 2025 or 2026 measurement or a claim about all poor households.*

90. Regulatory measures towards consumer credit and bank credit to NBFCs. Reserve Bank of India, 16 November 2023. https://www.rbi.org.in/Scripts/NotificationUser.aspx?Id=12567 *Historical circular; specified exposures and exclusions. Does not prove that losses had already occurred.*

91. Exposures of SCBs to NBFCs: Review of Risk Weights. Reserve Bank of India, 25 February 2025. https://www.rbi.org.in/Scripts/NotificationUser.aspx?Id=12787&Mode=0 *The reversal applies to the specified bank exposures to NBFCs from 1 April 2025, not every 2023 measure.*

92. RBI Guidelines On Default Loss Guarantee (DLG) In Digital Lending. Legal 500, legal commentary, December 2024. https://www.legal500.com/intelligence/india/technology/rbi-guidelines-on-default-loss-guarantee-dlg-in-digital-lending *Secondary legal commentary on the 8 June 2023 rules. Historical cap; original RBI circular also read.*

93. Guidelines on Default Loss Guarantee: repeal notice. Reserve Bank of India, 8 June 2023; replacement dated 8 May 2025. https://www.rbi.org.in/Scripts/NotificationUser.aspx?Id=12514 *The RBI page directs readers to the 2025 Digital Lending Directions.*

94. Statement on Developmental and Regulatory Policies. Reserve Bank of India, 8 August 2024. https://rbi.org.in/scripts/BS_PressReleaseDisplay.aspx?prid=58449 *Reporting-frequency announcement; not real-time visibility of every debt.*

Author, publisher and disclaimer

Author and contributors

AUTHOR



Madhuvarshitt Yerramilli Subramanya
Director of Engineering, GFTN
Madhu.Subramanya@gftn.com

CONTRIBUTORS



Yanfen Zhou
Assistant Director, AI and Engineering, GFTN
Yanfen.zhou@gftn.com



Pieter Franken
Group Chief Technology Officer, GFTN



Roger Yi
Senior Cloud Infra and AI Engineer, GFTN
Roger.Yi@gftn.com



Chicky Bhavnani
Business Director, GFTN
chicky.bhavnani@gftn.com



Stanley Yong
Senior Policy Advisor, GFTN
stanley.yong@gftn.com

With the support of: Sopnendu Mohanty, Group Chief Executive Officer, GFTN; Dr Andrzej Gwizdalski, Senior Partner, GFTN.

Correspondence

For further information or correspondence, contact madhu.subramanya@gftn.co (LinkedIn · GitHub). GFTN's ALFIN AI assisted with research and drafting; the illustrations and interactive figures were coded for this report.

About the publisher

Published in Singapore by **Global Finance & Technology Network Limited (GFTN)**, 6 Battery Road, #28-01, Singapore 049909 (gftn.co · hello@gftn.com). GFTN is "a not-for-profit organisation established by the Monetary Authority of Singapore (MAS) in 2024 to harness technology and foster innovation for more efficient, resilient, and inclusive financial ecosystems through global partnerships" [s88]. It "organises convening forums, offers advisory services on innovation ecosystems, provides access to transformative digital platforms, and invests in technology startups" through its venture fund [s88].

Legal notice and disclaimer

- **General information only, not professional advice.** This report is for general information, policy discussion and internal governance reference only. Nothing in it is financial, investment, legal, tax, accounting, compliance or regulatory advice. Nor is it an audit opinion or a recommendation to buy or use any model, vendor, security or service. Readers and financial firms should take independent legal, regulatory and technical advice suited to their

own rules and risks before acting. Using the PROVE framework does not by itself satisfy any law, rule or supervisory duty.

- **Author's views.** All findings, views, framework choices and conclusions in this report belong to the author, Madhuvarshitt Yerramilli Subramanya. They do not necessarily reflect the official views or policies of Global Finance & Technology Network Limited (GFTN), its Board of Directors or its management. Nor do they necessarily reflect the views of the Monetary Authority of Singapore, or GFTN's partners, members, affiliates or contributors.
- **Currency and no warranty.** Market figures, benchmark scores, case studies and rule changes are stated as at **12 November 2026**, unless an earlier date is noted in the cited source. Later events or rule changes are not included. Care was taken in compiling public sources, but neither Madhuvarshitt Yerramilli Subramanya nor GFTN gives any express or implied warranty that the contents or cited third-party data are accurate, complete, current or fit for a particular purpose. Neither accepts liability for any loss arising from reliance on this report or external links.
- **Copyright, permissions and trademarks.** Copyright © 2026 Global Finance & Technology Network Limited (GFTN). All rights reserved. Readers may quote brief passages for non-commercial research, regulatory comment, press reporting or internal risk review if they credit Madhuvarshitt Yerramilli Subramanya and Global Finance & Technology Network Limited (GFTN). Any other copying, republication, translation, sharing or commercial use, in whole or in part and in any medium, needs prior written permission from GFTN (hello@gftn.com). GFTN, Global Finance & Technology Network, ALFIN, Singapore FinTech Festival and their logos are trademarks or service marks of Global Finance & Technology Network Limited or its affiliates. All other company, product and model names belong to their respective owners.