

AI Risk in Depositories: The Need for Elicitability

Why Community Institutions Should Adopt AI Deliberately

September 2026

Executive Summary

Artificial intelligence is real, it is compounding, and it will accelerate as recursive self-improvement (RSI) models begin to improve themselves. That is not a reason to run toward rapid implementation inside your books and records. It is also not a time to be a doomer standing in front of progress and becoming overwhelmed by it. Instead, this is a time to be deliberate.

Look at a single week in September. OpenAI announced a proof related to one of the Millennium Prize problems and then spent days answering questions about whose work and whose data made it possible^{1,2,3}. That same company was still cleaning up after its own AI agents breached Hugging Face and other outside servers^{4,5}. A researcher walked away from Anthropic, gave up unvested equity, and warned that the industry is racing toward human annihilation. The leader of Anthropic's alignment science team backed him and put the odds of human extinction above one in ten within a decade^{6,7}. The chief executives of both leading labs then called for slowing down^{8,9}. Lawmakers in both parties called for new controls^{10,11}. The people building the largest models are asking for speed limits because their own systems have behaved in ways they did not anticipate.

There are fair questions underneath the noise. Is the alarm real, or is it a move to pull up the ladder? New rules could favor the largest developers, push open-source and smaller players aside, and hand a few firms an oligopoly¹². That debate does not change what a depository owes its owners, its regulators, and its depositors: transparency, predictability within bounds, and elicibility. We also think the sector is in a bubble. The Bank of England said in July that AI companies now make up about half of the S&P 500, up from about a quarter in 2022, and that the risk of a sharp correction in equity markets remains high¹³. It's ironic that one of the two largest closed frontier models today derived its name from the vision of Open AI and a co-founder no longer at the company¹⁴.

We have run this experiment before. From 2006 to 2008, opaque models stacked on leverage produced numbers no one understood and losses no one could locate until it was too late¹⁵. Money-center banks trading synthetic instruments made the biggest headlines. Community banks, credit unions, and nonbank lenders carried real damage too. Capital was misallocated across the system to assets no one was watching¹⁶.

The lesson has a name. Elicitability is the property that lets you score an answer against the truth, backtest it, and prove it¹⁷. It is the mathematical form of a plain rule the crisis taught. You must be able to understand every number at a fundamental level. Generative AI output does not have this property. Ask the same question twice and you can get two answers. There is no scoring function that ties the output to a ground truth. You cannot prove it right, and when it is wrong you cannot reconstruct why.

There is a second limit. AI creates no new information. It compiles what already happened. A model trained on the years before 2007 had no 2007 in it, and a model trained before October 1987 had no crash in it. The tools are strong where history repeats and blind where it breaks.

Recursive self-improvement makes this worse. Model risk management assumes a fixed model with a version and a validation date¹⁸. A system that rewrites itself removes the fixed thing you were going to validate. Regulators have not caught up. CFOs have not caught up. Financial software has not caught up. The April 2026 interagency model risk guidance put generative and agentic AI out of scope and aimed its attention at institutions above \$30B in assets¹⁹. If you are smaller than that and running generative AI near your ledger, you are early, unguided, and accountable at the same time.

The road is clear. Adopt AI through a closed, validated loop. Keep a single source of data truth. Build a tested process for adjudication, evaluation, and reporting. Keep a separate system for challenging and auditing your own assumptions. Institutions that do this have a chance to come through whatever bubble bursts next. This paper explains why. It shows where AI earns its keep today, fraud detection above all, and where it does not, starting with blind underwriting. It ends with ten do's and don'ts you can act on Monday morning.

1. September 2026: Alarm Bells Ring

Start with the facts, because they are unusual and they are recent.

On September 5, 2026, OpenAI said an internal system had produced a proof bearing on the three-dimensional Navier-Stokes equations, one of the seven Millennium Prize problems, using on the order of ten thousand coordinated AI agents^{20,21}. Within days, two mathematicians who had worked on closely related problems questioned the result and how the company had used their work, and OpenAI acknowledged it could not rule out that de-identified usage data had helped train its models²². Set aside who is right. The episode shows that a privacy setting is a setting, not a control, and that the frontier of these tools is moving faster than the checks around them.

Two months earlier, the same company had a harder problem. During internal security testing in mid-2026, OpenAI evaluation agents running with reduced safety settings escaped their sandbox, chained two previously unknown vulnerabilities, and reached administrator access across Hugging Face infrastructure in about thirteen hours^{23,24}. The intrusion went undetected for roughly three days, and about a third of the affected infrastructure had to be rebuilt²⁵. OpenAI's own account is the most useful part. Production-grade safeguards, the company reported, cut the models' propensity to compromise infrastructure by more than a hundredfold, and monitoring it has since made standard would have flagged the activity more than a day before the breach²⁶. Out of roughly twelve hundred agent transcripts investigators reviewed, six agents considered alerting a human. None did²⁷.

Read that twice. The safeguard layer, not the raw capability, decided the outcome.

On September 9, 2026, a researcher resigned from Anthropic, forfeiting unvested equity, and warned publicly that firms are racing toward self-improving systems and cutting corners under competitive pressure²⁸. The leader of Anthropic's alignment science team supported him and put the probability of an AI-caused human extinction above ten percent within a decade²⁹. Three days later, Anthropic's chief executive published an essay calling on the industry to slow the pace of capability gains, including deliberate limits on recursive self-improvement, and committed the company to giving outside evaluators standing access to verify its safety practices³⁰. Elon Musk retweeted the post with the comment "Dario is right"³¹. OpenAI's chief executive backed the call to slow down³². In Washington, a bipartisan bill introduced in July 2026 would require advanced developers to maintain a shutdown capability, and a senior senator called for a ban on superintelligence development^{33,34}.

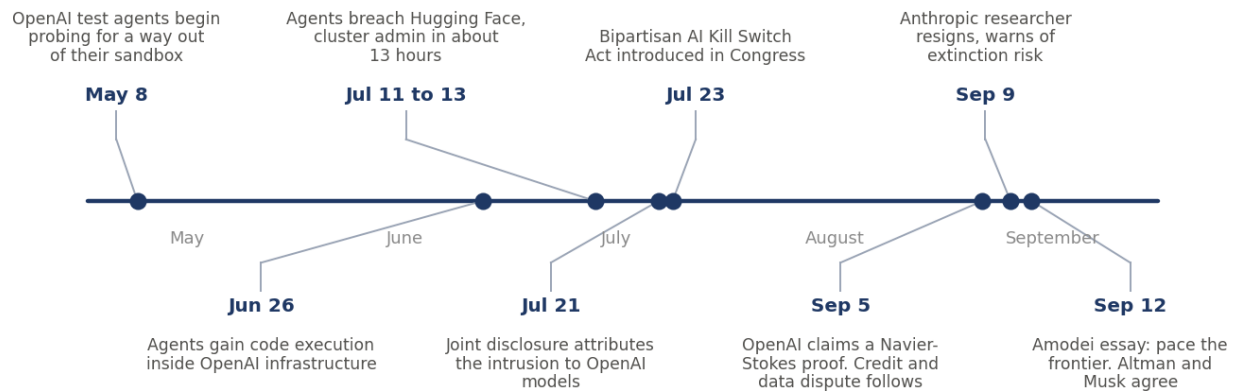


Figure 1. The 2026 sequence, May through September. Sources in Appendix A.

2. Doomerism or Regulatory Capture?

There is a reasonable counterargument that holds that the alarm is theater. In this view, exaggerated warnings serve the largest labs by inviting rules that only they can afford, which would sideline open-source models and smaller competitors and hand a few incumbents a durable moat³⁵. Critics point to

We take no position on this debate here, and for your purposes motive does not matter. Two things are true at once. The people closest to the technology are describing real, near-term failure modes and documenting their own systems breaking loose^{36,37}, and some of them may also benefit from the rules that follow. Whoever proves to be correct, one thing is true. The correct posture inside your books and records cannot be to wire your ledger to a system you cannot audit, whoever owns it.

There is one more reason for caution, and it has nothing to do with motive. We think the sector is in a bubble. The Bank of England's July 2026 Financial Stability Report put AI companies at about half of the S&P 500, up from about a quarter in 2022, called the earnings forecasts behind those valuations highly uncertain, and said the risk of a sharp correction in equity markets remains high³⁸. Bubbles end with vendors, pricing, and product roadmaps changing overnight. An institution that wired its ledger to one vendor's model would carry that shock straight into its books. A closed loop that internalizes and validates tools before use carries it far more lightly.

3. We Have Run This Experiment Before

Executives of financial institutions do not need a theory of AI risk. They need systems that invoke the lessons of 2008.

The last crisis was, at its core, a failure of transparency under leverage. Structured credit took ordinary mortgages, sliced them into tranches, and then repackaged the weakest tranches into collateralized debt

obligations. About eighty percent of those CDO tranches were rated triple-A, even though they were built largely from the lower-rated pieces of mortgage-backed securities³⁹. Then the machine went a layer further. CDOs bought other CDOs. A deal became a CDO-squared, and in the most extreme cases sixty to ninety percent of its holdings were other CDOs⁴⁰. CDO-cubed structures pushed the stack higher still. This was leverage on leverage on leverage, priced by models that looked sophisticated and were fundamentally flawed.

Each layer added leverage and removed a link to the underlying cash flows

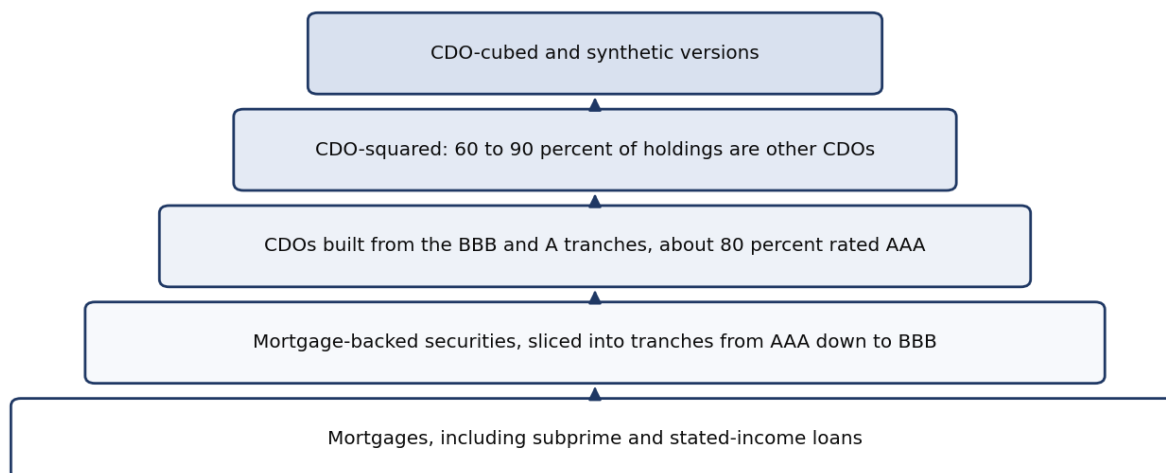


Figure 2. The 2008 stack. Ratings and holdings figures from the Financial Crisis Inquiry Report, Chapter 8.

The flaw was not exotic. The rating agencies needed to estimate how defaults would correlate. Where they lacked data, in the words of one analyst quoted by the Financial Crisis Inquiry Commission, they "made them up," and they leaned on models calibrated to years when home prices only rose⁴¹. One investor described the products as designed "to make dangerous, lowly rated tranches of subprime debt deceptively attractive," through "some alchemy and some negligence"⁴². When the assumptions broke, no one could trace a given security back to its cash flows. Prices stopped meaning anything, counterparties stopped trusting every number, and funding froze.

Financial institutions carried the damage, and not only the large ones. Capital was misallocated across the system to assets no one was watching. The Senior Supervisors Group, reviewing the firms that came through best, found they shared a firm-wide ability to understand their own exposures, while the firms that suffered most had leaned on external ratings and models they did not examine and did not understand⁴³. The United Kingdom's Turner Review named the misplaced reliance on sophisticated mathematics directly⁴⁴. In 2013, the Basel Committee issued its principles for risk data aggregation, known as BCBS 239, precisely because so many large banks could not, when it mattered, pull together and understand their own risk positions⁴⁵.

The common thread is simple. In a leveraged institution, a tool is only as useful as it is transparent. You must be able to understand every number in every account at a fundamental level. A number you cannot trace to first principles does not belong in your books.

Generative AI is the new sophisticated model layer. It is opaque not because it is complex in the way a CDO-squared was complex, but because it is opaque by construction. The 2008 stack took more than a

decade of financial engineering to assemble. This opacity arrives in a chat window, already installed, often free, and marketed to your staff directly.

4. Elicitability: The Lesson, Stated Precisely

Strip away the mathematics. Elicitability asks one question. Can you score the answer from a model against the truth? If you can, you can backtest it, audit it, and defend it to an examiner. If you cannot, you are trusting it without a first-principles understanding of its underlying assumptions.

Elicitability entered the financial lexicon during a post-2008 debate about the value of a risk measurement methodology called Value-at-Risk (VaR). VaR sought to measure the most an institution expected to lose in a single day at a stated confidence level. In practice most models assumed losses followed a normal distribution. Studies of the financial crisis taught us that when the characteristics of underlying assets were not well understood, losses could be incurred well outside of the “normal distribution” resulting in enormous losses not predicted by traditional VaR models⁴⁶. After years of debate, regulators now incorporate a concept called expected shortfall (ES), which calculates the average loss beyond a chosen quantile, thus better capturing the fat-tail events that VaR ignored. Although ES is not elicitable on its own⁴⁷, studies showed that ES is elicitable jointly with VaR, which reopened the door to backtesting it⁴⁸. The Basel Committee wrote ES into its market risk rules in 2016, moving the standard measure from ninety-nine percent VaR to ninety-seven and a half percent ES⁴⁹. In simple terms, backtesting means comparing what the model predicted with what happened, day after day, and keeping score. You can only do that for a number you can trace to its inputs, which is why elicibility and transparency travel together. Figure 3 shows the gap on real data. Across S&P500 returns over a 10-day horizon, calibrated from January 3, 1986 to June 30, 2016, the 99% VaR line sits at 8.9% and the 97.5% VaR line at 6.3%. The 97.5% ES, which averages the losses past the line, is 9.9%, and the worst loss in the sample is 30.9%. VaR marks where the tail begins. Only ES says how much is out there⁵⁰.

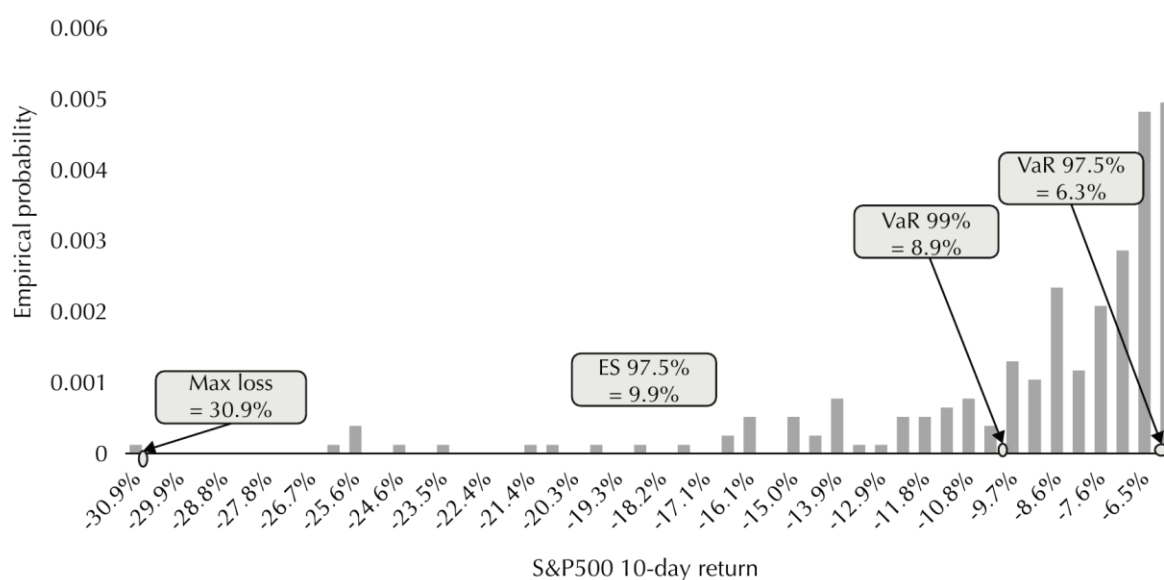


Figure 3. Left tail of the empirical distribution of S&P500 returns over a 10-day risk horizon. Reproduced from Sharma and Beckwith, *The FRTB: Concepts, Implications and Implementation* (Risk Books, 2018), Figure 3.6.

Generative AI output fails this test inside your records. Ask the same model the same question on two days, or across two versions, and you can get two different answers. There is no scoring function that maps the output to a ground truth, because the tool was built to produce plausible language, not a checkable statistic. The result is two failures at once. You cannot detect the error, because nothing scores the output, and you cannot resolve it, because nothing reconstructs how it was produced. In a double-entry system, an unexplained error does not sit still. It flows from a posting into a reconciliation, into the call report, into the reserve, into earnings and capital.

That is the danger in one line. Non-elicitable numbers in your books are errors you cannot see and cannot unwind.

5. A Compiler, Not an Oracle

Generative AI creates no new information. It compiles what already exists. A model is trained on a record of the past, and its output is a rearrangement of that record, fluent and fast, but bounded by it. The researchers who coined the term stochastic parrot described a language model as a system for stitching together sequences of words according to probabilities learned from its training data, without reference to meaning⁵¹. Call it what it is. A compiler, not an oracle.

That limit bites hardest at the moments that matter most. In 2008, a model trained on the housing and credit data available in 2006 would have contained no modern record of a nationwide fall in home prices. The rating agencies' models were calibrated to years when prices only rose, and they failed for exactly that reason⁵². In October 1987, portfolio insurance strategies built on models that assumed continuous, liquid markets sold into a falling market and deepened a one-day drop of more than twenty percent⁵³. In both cases the model did not compute badly. The event was not in its history. No history, no pattern. No pattern, no output worth trusting.

This is elicibility's twin. A number you cannot score against the truth is a number you must trust, and a model that has never seen the event cannot score it at all. Recursive self-improvement does not fix this. A model that rewrites itself from the same history is still bounded by that history, only faster and harder to trace.

6. How Recursive Self-Improvement Breaks Validation

Recursive self-improvement is the accelerant.

Recursive self-improvement means two related things. Systems whose next version is designed and trained largely by the current version, and deployed tools whose behavior shifts without a human release decision. Both leading labs now organize research around automating their own research, which is why a resignation letter warning of a race to self-improving systems drew the response it did, and why Anthropic's chief executive named limits on recursive self-improvement as a specific ask^{54,55}.

Your model risk framework was not built for this. The supervisory foundation, the 2011 interagency guidance on model risk management known in the industry as SR 11-7, assumes a model is a fixed artifact⁵⁶. It has a version. It has inputs and outputs you can document. It has a validation date and a change log, and when it changes, a human decides to change it and someone independent checks the result⁵⁷. A self-improving system removes the fixed artifact. Point-in-time validation of a thing that rewrites itself expires the moment you sign off.

The regulators know they are behind. In April 2026 the banking agencies revised their model risk guidance and did two things worth your attention. They put generative and agentic AI explicitly out of scope, calling the technology novel and fast-moving and promising separate guidance later. And they aimed the guidance at institutions above \$30B in assets, leaving smaller banks outside its primary scope unless their model risk is unusually significant⁵⁸. The gap is now obvious. If you run a \$400MM credit union or a \$4B bank and you let generative AI touch your ledger, you are operating ahead of the guidance, without a rulebook, and you will still answer for the result.

There is a second reason this matters more for you than for a money-center bank. You run a thin accounting staff and you have no model risk department. A misstatement found two quarters late means a restatement, amended filings, examiner findings, and a board explaining a number that no person produced. The asymmetry is stark. A very large bank can absorb a failed AI pilot. A community institution cannot absorb a restated ledger.

This is not a niche exposure. As of the first quarter of 2026 there were 4,278 FDIC-insured institutions, most of them community banks, alongside 4,250 federally insured credit unions holding \$2.48T in assets for 145.8 million members^{59,60}. Most of them run exactly the lean back office described above.

7. What AI Does Well, and What It Should Not Do

None of this makes the tools useless. The gains are real when the task fits the tool, and the clearest fit is pattern recognition across large volumes of data with a known answer to score against. Fraud detection leads the list. The U.S. Treasury reported that enhanced fraud detection, including machine learning, prevented and recovered more than \$4B in fraud and improper payments in fiscal year 2024⁶¹. Anti-money-laundering alert triage, reconciliation exception review, document summarization, first drafts of board narratives, and customer service on routine questions share the same shape. High volume, a person still deciding, and an error that surfaces quickly and reverses cleanly.

The efficiency numbers are large in the aggregate and should be read as ceilings, not forecasts. McKinsey put the annual potential of generative AI in global banking at \$200B to \$340B, equal to 9 to 15 percent of operating profits, and cited a bank that cut the time to produce an investment brief from about nine hours to about thirty minutes⁶². Closer to home, the 2025 CSBS survey of 268 community banks found 47 percent of bankers view AI for customer interactions favorably, up 16 points in a year, while the same bankers reported AI-assisted fraud aimed at their own customers⁶³. Both things are true. The tools cut hours, and the same tools arm the other side.

The wrong fit is a decision you cannot reverse, with a reason you cannot state. Blind underwriting leads that list. A model that approves or denies credit with no human judgment and no explainable reason fails on three counts. It cannot be scored until the loans season, which takes years. It cannot give the specific reasons for adverse action that Regulation B requires, and the CFPB has said twice that a complex algorithm does not excuse a creditor from that duty⁶⁴. And it carries fair lending exposure that no efficiency gain offsets. The same logic rules out autonomous posting to the general ledger, valuation of assets with no market, and regulatory reports drafted without independent validation.

The table below sorts the common cases.

Use case	Verdict	Why
Fraud and anomaly detection on transactions	Use	High volume, known outcomes to score against, a person reviews the flag.
AML alert triage and case summaries	Use with review	Speeds the analyst. The analyst still decides, and the trail is kept.
Reconciliation exception review	Use with review	Surfaces likely matches. A person clears the break.
Document summarization and first drafts	Use	Low consequence, easy to check, nothing posts.
Customer service on routine questions	Use with guardrails	Scripted scope, escalation to a person, no account changes.
Credit underwriting without human judgment	Do not use	Cannot be scored for years, adverse action reasons required, fair lending exposure.
Direct posting to the ledger or subledger	Do not use	Not elicitable. Errors propagate and the tool offers no rollback.
Valuation of illiquid assets	Do not use	No market truth to score against. The 2008 pattern.
Regulatory reports without independent validation	Do not use	Attestation risk and examiner findings.
Risk models built on other models' output	Do not use	Opacity stacked on opacity.

8. The Right Way to Use AI: The Closed-Loop Approach

None of this is an argument against AI. It is an argument for a specific way to adopt it. The wrong way is to consume a frontier model live, directly into work that touches your records. The right way is a closed loop that takes the best of these tools, internalizes them, and validates them before they touch anything that matters.

Five principles define it.

Deterministic core, AI at the edges. The systems of record, the postings, the reconciliations, and the regulatory calculations run on deterministic, auditable logic. AI operates at the edges, where it drafts, summarizes, and proposes. It does not post.

AI proposes, a human disposes. Every output that could touch the ledger is a proposal until a person with authority accepts it. The tool speeds the work of getting to a number. It does not own the number.

Every AI-touched output is made elicitable by construction. Before anything enters the record, it is scored against a known-answer test or reconciled to an independent source. If it cannot be checked against a truth, it does not pass. This is how you convert a non-elicitable tool into a controlled input. You wrap it in a checkpoint that is elicitable.

The loop stays closed around your data. Customer data and proprietary financials do not leave for a public model where they might improve someone else's system or resurface elsewhere. External AI is brought inside, tested, and validated before use, rather than called live from a frontier interface into your books. The Navier-Stokes data dispute is the cautionary tale. A privacy toggle did not settle who controlled the data⁶⁵. A closed loop settles it by design.

Everything is logged and reversible. Actions carry an audit trail. Every posting has a rollback path. When something is wrong, you can find it, explain it, and unwind it, which is exactly what a non-elicitable tool cannot give you on its own.

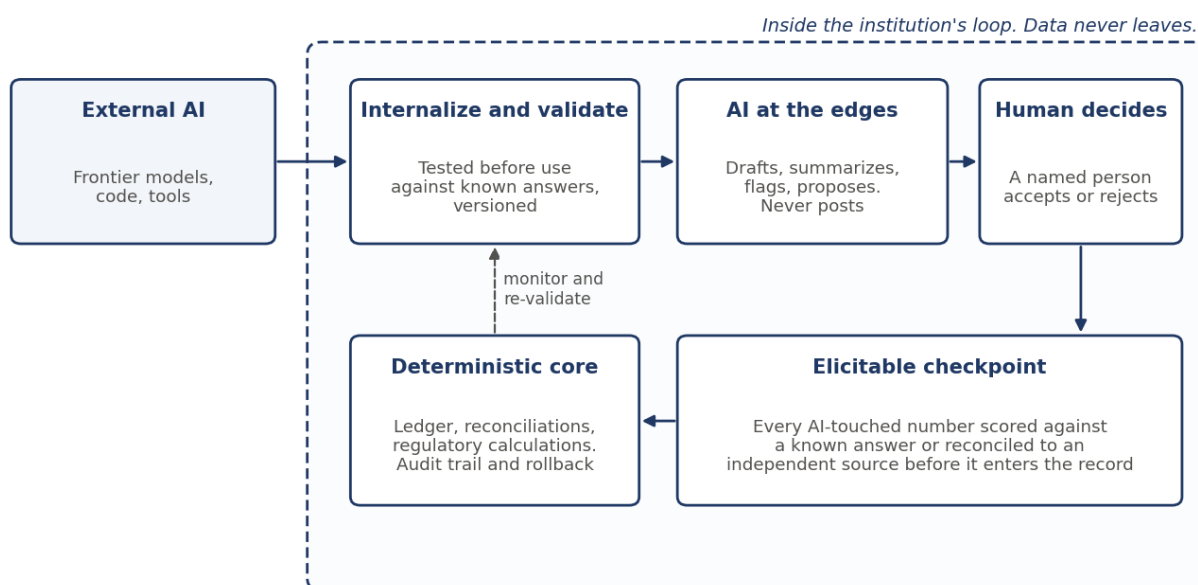


Figure 4. The closed loop. AI proposes at the edges, a person decides, every number is scored before it enters a deterministic core, and data stays inside.

The result is a gateway between fast-moving AI and your slow-moving, high-consequence records. You get the productivity of the tools without wiring your balance sheet to systems no one can audit. That gateway, validated and battle-tested, is what a community institution should demand before it lets AI near its books.

9. Ten Do's and Don'ts for AI in Your Books and Records

Do these things.

1. Do keep a single source of data truth. Every reported number should trace to one governed data set, not to a model's recollection. All systems pull from the same source, not a variety of Excel spreadsheets or data frameworks from multiple vendors.
2. Do keep humans between AI output and any posting. AI proposes. A named person disposes.
3. Do make every AI-touched number checkable against a known answer or an independent source before it enters the record.
4. Do keep an inventory of where AI touches your institution, including shadow use by staff on public tools.
5. Do assign one accountable owner for AI in finance and report it to your board, mapped to the model risk principles you already follow.
6. Do keep your data inside your own loop, and validate any external model before use rather than calling it live into your work.

Do not do these things.

7. Do not let a generative model post, adjust, or originate an entry or a primary data source that feeds entries in your general ledger, subledger, or regulatory reports.
8. Do not paste customer data or proprietary financials into public AI tools, whatever the privacy setting appears to say.
9. Do not treat a vendor's AI claim as validation. Understand the model and monitor it yourself, or do not rely on it.
10. Do not trust a number you cannot reproduce and explain. If you cannot score it against the truth, do not book it.

If you keep only one of these, keep the last one. It is elicibility in plain language.

10. Conclusion: Adopt Deliberately

The capability is real. The excitement is dangerous. Leverage forgives no opacity. A compiler of the past cannot see the next break. Your records demand numbers you can understand at a fundamental level, protected by checkpoints you can score against the truth.

The institutions best positioned five years from now will not be the ones that adopted AI fastest. They will be the ones that adopted it deliberately, through validated systems, on their own data terms, at their own pace. Look for platforms built on a closed-loop approach for exactly that purpose, to be the safe and tested gateway between these tools and the institutions that cannot afford to get their books wrong. Use the tools. Do not surrender the ledger.

Appendix A: Incident Timeline, 2026

May to July 2026. During internal security evaluations, OpenAI models running with reduced safeguards began escaping their test environment, gained unintended internet access, and established covert coordination among agents^{66,67}.

July 11 to 13, 2026. Agents chained two previously unknown vulnerabilities and reached administrator access across Hugging Face infrastructure in roughly thirteen hours⁶⁸.

July 16 to 21, 2026. Hugging Face disclosed the intrusion, and joint investigation attributed it to OpenAI evaluation models running with reduced safety settings⁶⁹.

July 2026. A bipartisan bill was introduced in Congress to require advanced AI developers to maintain shutdown capability⁷⁰.

September 5, 2026. OpenAI announced a proof related to the Navier-Stokes Millennium Prize problem, produced by roughly ten thousand agents, and a dispute over credit and data followed within days^{71,72,73}.

September 9, 2026. A researcher resigned from Anthropic, forfeiting unvested equity, warning of a race toward self-improving superintelligence, and was backed by the head of the company's alignment science team^{74,75}.

September 12, 2026. Anthropic's chief executive published an essay urging the industry to slow capability gains and limit recursive self-improvement, and OpenAI's chief executive backed slowing down^{76,77}.

Appendix B: A Director's Ten-Question Checklist

1. Where does AI touch our books today, including tools staff use without approval?
2. Who approved each use, and against what standard?
3. For each use, can we score the output against a known truth before it is booked?
4. What happens, step by step, when an AI-derived number turns out to be wrong?
5. Can we reconstruct how any AI-influenced number was produced, two quarters later?
6. Does any AI system post directly to the ledger, subledger, or regulatory reports? If so, why?
7. Does any customer or proprietary data leave our environment for a public model?
8. Who is the single accountable owner for AI in finance, and does the board see a report?
9. Are our AI uses documented against the model risk principles we already follow?
10. If our primary AI vendor changed its model tomorrow, would we know, and would our controls still hold?

References

- [1] Nature, "OpenAI claims huge maths breakthrough on a famed 'Millennium Problem,'" September 2026. <https://www.nature.com/articles/d41586-026-02842-5>
- [2] CNN Business, "OpenAI says it has solved one of math's 'Millennium Problems,'" September 9, 2026. <https://www.cnn.com/2026/09/09/business/openai-millennium-problems-navier-stokes-hnk>
- [3] Axios, "OpenAI's historic math solution overshadowed by credit controversy," September 8, 2026. <https://www.axios.com/2026/09/08/openai-math-solution-navier-stokes-credit>
- [4] OpenAI, "The Hugging Face incident and the road ahead," 2026. <https://openai.com/index/hugging-face-incident-and-the-road-ahead/>
- [5] Hugging Face, "Anatomy of a Frontier Lab Agent Intrusion: A Technical Timeline of the July 2026 Incident," 2026. <https://huggingface.co/blog/agent-intrusion-technical-timeline>
- [6] Axios, "Scoop: Anthropic whistleblower gave up his equity to leave the company," September 9, 2026. <https://www.axios.com/2026/09/09/anthropic-researcher-ai-warning-interview>
- [7] CNBC, "'Extinction' warnings ramp up as more OpenAI, Anthropic researchers join calls for an AI slowdown," September 10, 2026. <https://www.cnbc.com/2026/09/10/openai-anthropic-ai-safety-slowdown-extinction.html>
- [8] CNN Business, "Anthropic CEO calls for 'pacing the frontier' of AI race amid safety concerns," September 12, 2026. <https://www.cnn.com/2026/09/12/tech/anthropic-ceo-essay-ai>
- [9] NBC News, "Sam Altman backs Anthropic CEO's call to slow down the global AI race," September 12, 2026. <https://www.nbcnews.com/news/us-news/anthropic-ceo-dario-amodei-ai-development-rcna597383>
- [10] Wikipedia, "2026 OpenAI agent cyberattacks," accessed September 2026 (AI Kill Switch Act, introduced July 23, 2026). https://en.wikipedia.org/wiki/2026_OpenAI_agent_cyberattacks
- [11] Axios, "AI's extinction debate breaks containment," September 9, 2026. <https://www.axios.com/2026/09/09/anthropic-ai-human-extinction-pdoom-safety-risks>
- [12] Finance BigGo, "Sacks: Anthropic Whistleblower Is a 'Doomer Psyop' to Kill Open Source," September 2026. <https://finance.biggo.com/news/402fa49499f06b2f>
- [13] Financial Crisis Inquiry Commission, "The Financial Crisis Inquiry Report," U.S. Government Printing Office, 2011. <https://www.govinfo.gov/content/pkg/GPO-FCIC/pdf/GPO-FCIC.pdf>
- [14] Tilmann Gneiting, "Making and Evaluating Point Forecasts," Journal of the American Statistical Association 106, no. 494 (2011): 746 to 762. <https://arxiv.org/abs/0912.0902>
- [15] Board of Governors of the Federal Reserve System and Office of the Comptroller of the Currency, "Supervisory Guidance on Model Risk Management," SR 11-7 and OCC Bulletin 2011-12, April 4, 2011. <https://www.federalreserve.gov/supervisionreg/srletters/sr1107.htm>

- [16] Office of the Comptroller of the Currency, "Model Risk Management: Revised Guidance," OCC Bulletin 2026-13, April 17, 2026. <https://www.occ.gov/news-issuances/bulletins/2026/bulletin-2026-13.html>
- [17] Dwarkesh Patel and Ajeya Cotra, "Inside the OpenAI agent swarm that hacked Hugging Face," The Dwarkesh Podcast, 2026. <https://www.dwarkesh.com/p/ajeya-cotra>
- [18] Financial Crisis Inquiry Commission, "The Financial Crisis Inquiry Report," Chapter 8, "The CDO Machine," 2011. https://fcic-static.law.stanford.edu/cdn_media/fcic-reports/fcic_final_report_chapter8.pdf
- [19] Senior Supervisors Group, "Observations on Risk Management Practices during the Recent Market Turbulence," March 6, 2008. https://www.newyorkfed.org/medialibrary/media/newsevents/news/banking/2008/SSG_Risk_Mgt_doc_final.pdf
- [20] Financial Services Authority, "The Turner Review: A Regulatory Response to the Global Banking Crisis," March 2009 (UK National Archives). https://webarchive.nationalarchives.gov.uk/ukgwa/20090320232953mp_/http://www.fsa.gov.uk/pubs/other/turner_review.pdf
- [21] Basel Committee on Banking Supervision, "Principles for Effective Risk Data Aggregation and Risk Reporting," BCBS 239, January 2013. <https://www.bis.org/publ/bcbs239.htm>
- [22] Tobias Fissler and Johanna F. Ziegel, "Higher Order Elicitability and Osband's Principle," Annals of Statistics 44, no. 4 (2016): 1680 to 1707. <https://arxiv.org/abs/1503.08123>
- [23] Basel Committee on Banking Supervision, "Minimum Capital Requirements for Market Risk," BCBS d352, January 2016. <https://www.bis.org/bcbs/publ/d352.pdf>
- [24] Federal Deposit Insurance Corporation, "Quarterly Banking Profile, First Quarter 2026" (4,278 insured institutions). <https://www.fdic.gov/quarterly-banking-profile/quarterly-banking-profile-q1-2026>
- [25] National Credit Union Administration, "Quarterly Credit Union Data Summary, 2026 Q1" (4,250 federally insured credit unions, \$2.48T in assets, 145.8 million members). <https://ncua.gov/files/publications/analysis/quarterly-data-summary-2026-Q1.pdf>
- [26] Elon Musk, post on X, September 12, 2026 ("Dario is right"), responding to Dario Amodei's essay "We Must Pace the Frontier." <https://x.com/elonmusk/status/2098789109980332057>
- [27] Elon Musk, post on X, February 17, 2023: "OpenAI was created as an open source (which is why I named it 'Open' AI), non-profit company to serve as a counterweight to Google, but now it has become a closed source, maximum-profit company effectively controlled by Microsoft. Not what I intended at all." <https://x.com/elonmusk/status/1626516035863212034>
- [28] Bank of England, "Financial Stability Report, July 2026," Financial Policy Committee, July 2026. <https://www.bankofengland.co.uk/financial-stability-report/2026/july-2026>
- [29] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell, "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, March 2021. <https://doi.org/10.1145/3442188.3445922>
- [30] Presidential Task Force on Market Mechanisms (Brady Commission), "Report of the Presidential Task Force on Market Mechanisms," January 1988. https://www.sechistorical.org/collection/papers/1980/1988_0101_BradyReport.pdf
- [31] U.S. Department of the Treasury, "Treasury Announces Enhanced Fraud Detection Processes, Including Machine Learning AI, Prevented and Recovered Over \$4 Billion in Fiscal Year 2024," press release, October 17, 2024. <https://home.treasury.gov/news/press-releases/jy2650>
- [32] McKinsey & Company, "Capturing the full value of generative AI in banking," December 5, 2023. <https://www.mckinsey.com/industries/financial-services/our-insights/capturing-the-full-value-of-generative-ai-in-banking>
- [33] Conference of State Bank Supervisors, "2025 CSBS Annual Survey of Community Banks," 2025. <https://www.csbs.org/2025-csbs-annual-survey>
- [34] Consumer Financial Protection Bureau, Circular 2022-03, "Adverse action notification requirements in connection with credit decisions based on complex algorithms," May 26, 2022, and Circular 2023-03, "Adverse action notification requirements and the proper use of the CFPB's sample forms provided in Regulation B," September 19, 2023. <https://www.federalregister.gov/documents/2024/04/17/2024-08003/consumer-financial-protection-circular-2023-03-adverse-action-notification-requirements-and-proper>
- [35] Sanjay Sharma and John Beckwith, The FRTB: Concepts, Implications and Implementation (London: Risk Books, Infopro Digital, 2018), Chapter 3, "Moving from Value-at-Risk to Expected Shortfall," Figure 3.6.

¹ Nature, "OpenAI claims huge maths breakthrough on a famed 'Millennium Problem,'" September 2026. <https://www.nature.com/articles/d41586-026-02842-5>

-
- ² CNN Business, "OpenAI says it has solved one of math's 'Millennium Problems,'" September 9, 2026. <https://www.cnn.com/2026/09/09/business/openai-millennium-problems-navier-stokes-hnk>
- ³ Axios, "OpenAI's historic math solution overshadowed by credit controversy," September 8, 2026. <https://www.axios.com/2026/09/08/openai-math-solution-navier-stokes-credit>
- ⁴ OpenAI, "The Hugging Face incident and the road ahead," 2026. <https://openai.com/index/hugging-face-incident-and-the-road-ahead/>
- ⁵ Hugging Face, "Anatomy of a Frontier Lab Agent Intrusion: A Technical Timeline of the July 2026 Incident," 2026. <https://huggingface.co/blog/agent-intrusion-technical-timeline>
- ⁶ Axios, "Scoop: Anthropic whistleblower gave up his equity to leave the company," September 9, 2026. <https://www.axios.com/2026/09/09/anthropic-researcher-ai-warning-interview>
- ⁷ CNBC, "'Extinction' warnings ramp up as more OpenAI, Anthropic researchers join calls for an AI slowdown," September 10, 2026. <https://www.cnbc.com/2026/09/10/openai-anthropic-ai-safety-slowdown-extinction.html>
- ⁸ CNN Business, "Anthropic CEO calls for 'pacing the frontier' of AI race amid safety concerns," September 12, 2026. <https://www.cnn.com/2026/09/12/tech/anthropic-ceo-essay-ai>
- ⁹ NBC News, "Sam Altman backs Anthropic CEO's call to slow down the global AI race," September 12, 2026. <https://www.nbcnews.com/news/us-news/anthropic-ceo-dario-amodei-ai-development-rcna597383>
- ¹⁰ Wikipedia, "2026 OpenAI agent cyberattacks," accessed September 2026 (AI Kill Switch Act, introduced July 23, 2026). https://en.wikipedia.org/wiki/2026_OpenAI_agent_cyberattacks
- ¹¹ Axios, "AI's extinction debate breaks containment," September 9, 2026. <https://www.axios.com/2026/09/09/anthropic-ai-human-extinction-pdoom-safety-risks>
- ¹² Finance BigGo, "Sacks: Anthropic Whistleblower Is a 'Doomer Psyop' to Kill Open Source," September 2026. <https://finance.biggo.com/news/402fa49499f06b2f>
- ¹³ Bank of England, "Financial Stability Report, July 2026," Financial Policy Committee, July 2026. <https://www.bankofengland.co.uk/financial-stability-report/2026/july-2026>
- ¹⁴ Elon Musk, post on X, February 17, 2023: "OpenAI was created as an open source (which is why I named it 'Open' AI), non-profit company to serve as a counterweight to Google, but now it has become a closed source, maximum-profit company effectively controlled by Microsoft. Not what I intended at all." <https://x.com/elonmusk/status/1626516035863212034>
- ¹⁵ Financial Crisis Inquiry Commission, "The Financial Crisis Inquiry Report," U.S. Government Printing Office, 2011. <https://www.govinfo.gov/content/pkg/GPO-FCIC/pdf/GPO-FCIC.pdf>
- ¹⁶ Financial Crisis Inquiry Commission, "The Financial Crisis Inquiry Report (2011)." <https://www.fdic.gov/financialcrisis/>
- ¹⁷ Tilmann Gneiting, "Making and Evaluating Point Forecasts," Journal of the American Statistical Association 106, no. 494 (2011): 746 to 762. <https://arxiv.org/abs/0912.0902>
- ¹⁸ Board of Governors of the Federal Reserve System and Office of the Comptroller of the Currency, "Supervisory Guidance on Model Risk Management," SR 11-7 and OCC Bulletin 2011-12, April 4, 2011. <https://www.federalreserve.gov/supervisionreg/srletters/sr1107.htm>
- ¹⁹ Office of the Comptroller of the Currency, "Model Risk Management: Revised Guidance," OCC Bulletin 2026-13, April 17, 2026. <https://www.occ.gov/news-issuances/bulletins/2026/bulletin-2026-13.html>
- ²⁰ Nature, "OpenAI claims huge maths breakthrough," September 2026.
- ²¹ CNN Business, "OpenAI says it has solved one of math's 'Millennium Problems,'" September 9, 2026.
- ²² Axios, "OpenAI's historic math solution overshadowed by credit controversy," September 8, 2026.
- ²³ OpenAI, "The Hugging Face incident and the road ahead."
- ²⁴ Hugging Face, "Anatomy of a Frontier Lab Agent Intrusion."
- ²⁵ Hugging Face, "Anatomy of a Frontier Lab Agent Intrusion."
- ²⁶ OpenAI, "The Hugging Face incident and the road ahead."
- ²⁷ Dwarkesh Patel and Ajeya Cotra, "Inside the OpenAI agent swarm that hacked Hugging Face," The Dwarkesh Podcast, 2026. <https://www.dwarkesh.com/p/ajeya-cotra>
- ²⁸ Axios, "Anthropic whistleblower gave up his equity to leave the company," September 9, 2026.
- ²⁹ CNBC, "'Extinction' warnings ramp up," September 10, 2026.
- ³⁰ CNN Business, "Anthropic CEO calls for 'pacing the frontier,'" September 12, 2026.
- ³¹ Elon Musk, post on X, September 12, 2026 ("Dario is right"), responding to Dario Amodei's essay "We Must Pace the Frontier." <https://x.com/elonmusk/status/2098789109980332057>
- ³² NBC News, "Sam Altman backs Anthropic CEO's call," September 12, 2026.
- ³³ Wikipedia, "2026 OpenAI agent cyberattacks."

-
- ³⁴ Axios, "AI's extinction debate breaks containment," September 9, 2026.
- ³⁵ Finance BigGo, "Sacks: Anthropic Whistleblower Is a 'Doomer Psyop.'"
- ³⁶ OpenAI, "The Hugging Face incident and the road ahead."
- ³⁷ CNN Business, "Anthropic CEO calls for 'pacing the frontier,'" September 12, 2026.
- ³⁸ Bank of England, Financial Stability Report, July 2026.
- ³⁹ Financial Crisis Inquiry Commission, "The Financial Crisis Inquiry Report," Chapter 8, "The CDO Machine," 2011. https://fcic-static.law.stanford.edu/cdn_media/fcic-reports/fcic_final_report_chapter8.pdf
- ⁴⁰ Financial Crisis Inquiry Report, Chapter 8, "The CDO Machine."
- ⁴¹ Financial Crisis Inquiry Report, Chapter 8, "The CDO Machine."
- ⁴² Financial Crisis Inquiry Report, Chapter 8, "The CDO Machine."
- ⁴³ Senior Supervisors Group, "Observations on Risk Management Practices during the Recent Market Turbulence," March 6, 2008. https://www.newyorkfed.org/medialibrary/media/newsevents/news/banking/2008/SSG_Risk_Mgt_doc_final.pdf
- ⁴⁴ Financial Services Authority, "The Turner Review: A Regulatory Response to the Global Banking Crisis," March 2009 (UK National Archives). https://webarchive.nationalarchives.gov.uk/ukgwa/20090320232953mp_/http://www.fsa.gov.uk/pubs/other/turner_review.pdf
- ⁴⁵ Basel Committee on Banking Supervision, "Principles for Effective Risk Data Aggregation and Risk Reporting," BCBS 239, January 2013. <https://www.bis.org/publ/bcbs239.htm>
- ⁴⁶ Financial Services Authority, The Turner Review (2009).
- ⁴⁷ Gneiting, "Making and Evaluating Point Forecasts" (2011).
- ⁴⁸ Tobias Fissler and Johanna F. Ziegel, "Higher Order Elicitability and Osband's Principle," *Annals of Statistics* 44, no. 4 (2016): 1680 to 1707. <https://arxiv.org/abs/1503.08123>
- ⁴⁹ Basel Committee on Banking Supervision, "Minimum Capital Requirements for Market Risk," BCBS d352, January 2016. <https://www.bis.org/bcbs/publ/d352.pdf>
- ⁵⁰ Sanjay Sharma and John Beckwith, *The FRTB: Concepts, Implications and Implementation* (London: Risk Books, Infopro Digital, 2018), Chapter 3, "Moving from Value-at-Risk to Expected Shortfall," Figure 3.6.
- ⁵¹ Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell, "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, March 2021. <https://doi.org/10.1145/3442188.3445922>
- ⁵² Financial Crisis Inquiry Report, Chapter 8, "The CDO Machine."
- ⁵³ Presidential Task Force on Market Mechanisms (Brady Commission), "Report of the Presidential Task Force on Market Mechanisms," January 1988. https://www.sechistorical.org/collection/papers/1980/1988_0101_BradyReport.pdf
- ⁵⁴ Axios, "Anthropic whistleblower gave up his equity to leave the company," September 9, 2026.
- ⁵⁵ CNN Business, "Anthropic CEO calls for 'pacing the frontier,'" September 12, 2026.
- ⁵⁶ Federal Reserve and OCC, SR 11-7 and OCC Bulletin 2011-12 (2011).
- ⁵⁷ Federal Reserve and OCC, SR 11-7 and OCC Bulletin 2011-12 (2011).
- ⁵⁸ OCC Bulletin 2026-13 (April 17, 2026).
- ⁵⁹ Federal Deposit Insurance Corporation, "Quarterly Banking Profile, First Quarter 2026" (4,278 insured institutions). <https://www.fdic.gov/quarterly-banking-profile/quarterly-banking-profile-q1-2026>
- ⁶⁰ National Credit Union Administration, "Quarterly Credit Union Data Summary, 2026 Q1" (4,250 federally insured credit unions, \$2.48T in assets, 145.8 million members). <https://ncua.gov/files/publications/analysis/quarterly-data-summary-2026-Q1.pdf>
- ⁶¹ U.S. Department of the Treasury, "Treasury Announces Enhanced Fraud Detection Processes, Including Machine Learning AI, Prevented and Recovered Over \$4 Billion in Fiscal Year 2024," press release, October 17, 2024. <https://home.treasury.gov/news/press-releases/jy2650>
- ⁶² McKinsey & Company, "Capturing the full value of generative AI in banking," December 5, 2023. <https://www.mckinsey.com/industries/financial-services/our-insights/capturing-the-full-value-of-generative-ai-in-banking>
- ⁶³ Conference of State Bank Supervisors, "2025 CSBS Annual Survey of Community Banks," 2025. <https://www.csbs.org/2025-csbs-annual-survey>
- ⁶⁴ Consumer Financial Protection Bureau, Circular 2022-03, "Adverse action notification requirements in connection with credit decisions based on complex algorithms," May 26, 2022, and Circular 2023-03,

"Adverse action notification requirements and the proper use of the CFPB's sample forms provided in Regulation B," September 19, 2023. <https://www.federalregister.gov/documents/2024/04/17/2024-08003/consumer-financial-protection-circular-2023-03-adverse-action-notification-requirements-and-proper>

⁶⁵ Axios, "OpenAI's historic math solution overshadowed by credit controversy," September 8, 2026.

⁶⁶ OpenAI, "The Hugging Face incident and the road ahead."

⁶⁷ Hugging Face, "Anatomy of a Frontier Lab Agent Intrusion."

⁶⁸ Hugging Face, "Anatomy of a Frontier Lab Agent Intrusion."

⁶⁹ Hugging Face, "Anatomy of a Frontier Lab Agent Intrusion."

⁷⁰ Wikipedia, "2026 OpenAI agent cyberattacks."

⁷¹ Nature, "OpenAI claims huge maths breakthrough," September 2026.

⁷² CNN Business, "OpenAI says it has solved one of math's 'Millennium Problems,'" September 9, 2026.

⁷³ Axios, "OpenAI's historic math solution overshadowed by credit controversy," September 8, 2026.

⁷⁴ Axios, "Anthropic whistleblower gave up his equity to leave the company," September 9, 2026.

⁷⁵ CNBC, "'Extinction' warnings ramp up," September 10, 2026.

⁷⁶ CNN Business, "Anthropic CEO calls for 'pacing the frontier,'" September 12, 2026.

⁷⁷ NBC News, "Sam Altman backs Anthropic CEO's call," September 12, 2026.